Maximum Likelihood Estimation in Gaussian AMP Chain Graph Models and Gaussian Ancestral Graph Models

Mathias Drton

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2004

Program Authorized to Offer Degree: Statistics

University of Washington

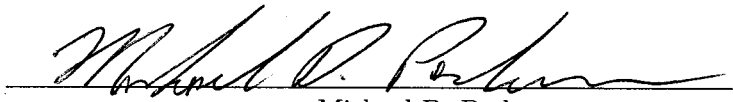
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Mathias Drton

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Co-Chairs of Supervisory Committee:

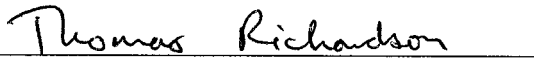


Michael D. Perlman

Thomas S. Richardson

Reading Committee:

Michael D. Perlman



Thomas S. Richardson

Steen A. Andersson

Date: 5/25/2004

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Nathias Ihen

Date 5/25/2004

University of Washington

Abstract

Maximum Likelihood Estimation in Gaussian AMP Chain Graph Models and Gaussian Ancestral Graph Models

by Mathias Drton

Co-Chairs of Supervisory Committee:

Professor Michael D. Perlman
Department of Statistics

Professor Thomas S. Richardson
Department of Statistics

Graphical Markov models use graphs to represent dependencies between stochastic variables. Via Markov properties, missing edges in the graph are translated into conditional independence statements, which, in conjunction with a distributional assumption, define a statistical model. This thesis considers maximum likelihood (ML) estimation of the parameters of two recently introduced classes of graphical Markov models in the case of continuous variables with a joint multivariate Gaussian distribution. The two new model classes are the AMP chain graph models, based on chain graphs equipped with a new Markov property, and the ancestral graph models, based on a new class of graphs. Both classes generalize the widely used models based on acyclic directed graphs (Bayesian networks) and undirected graphs (Markov random fields).

In this thesis, we first show that the likelihood of AMP chain graph and ancestral graph models may be multimodal. Next, we combine existing techniques (iterative proportional fitting, generalized least squares) into an algorithm for ML estimation in AMP chain graph models. For the ancestral graphs, we develop an ML estimation algorithm based on a new iterative conditional fitting (ICF) idea, which in the considered Gaussian case can be imple-

mented using least squares regression on synthetic variables. We derive the ICF algorithm in the special case of bidirected graphs, also termed covariance graphs, and subsequently generalize it to cover arbitrary ancestral graphs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisors Michael Perlman and Thomas Richardson for introducing me to graphical models and sharing with me the research questions that ultimately led to this thesis. Thanks to them also for financially supporting me with the NSF grants DMS 99-72008 and DMS 00-71818, giving me opportunities to attend conferences and workshops, and helping me plan my future career. I also wish to thank my committee member Steen Andersson, in particular for telling me about his research in graphical models when I first met him during his visit to Augsburg. My thanks go to Steffen Lauritzen for his helpful comments at the IMS/ISBA conference, inviting me to the workshop in Denmark, and writing me letters of recommendation. Finally, I would like to thank Friedrich Pukelsheim whose teaching provided me with a solid foundation in mathematical statistics and on whose help and support I could count on from my first semesters as an undergraduate at the Universität Augsburg until now.

Chapter 1

INTRODUCTION

1.1 Graphical models

Graphical Markov models, or simply graphical models, use graphs to represent dependencies between stochastic variables. For a given graph, so-called Markov properties yield a list of conditional independence relations that, in conjunction with distributional assumptions, are used to define a statistical model. Monographs providing an overview of the field of graphical models are, for example, Pearl (1988), Whittaker (1990), Andersen et al. (1995), Cox and Wermuth (1996), Lauritzen (1996) and Edwards (2000).

Most commonly, graphical models are based either on undirected graphs, in which case the models are also referred to as Markov random fields, or on acyclic directed graphs (ADG or more commonly DAG), in which case the models are also called Bayesian networks. An undirected graph and a DAG are shown in Figure 1.1(a) and (b), respectively. In order to provide more general independence models, chain graphs, as shown in Figure 1.1(c), were introduced by Lauritzen and Wermuth (1989), Wermuth and Lauritzen (1990) and Frydenberg (1990). Their Markov property for chain graphs is also called the LWF Markov property.

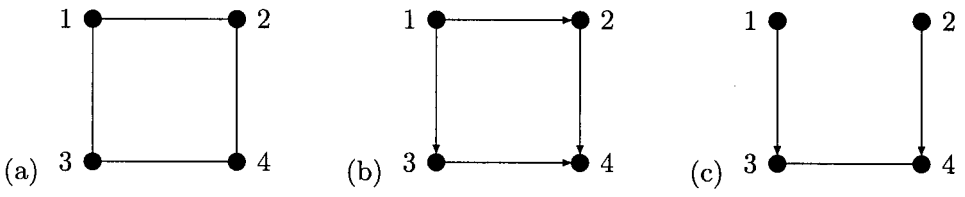


Figure 1.1: (a) An undirected graph, (b) an acyclic directed graph, and (c) a chain graph.

The distributional assumptions combined with the different types of graphs are the following. If all variables are discrete, then graphical models are based on multinomial or Poisson sampling and fall in the framework of contingency tables (see e.g. Lauritzen, 1996, §4). If all variables are continuous, then it is commonly assumed that they follow a joint multivariate Gaussian = normal distribution (for an exception see Capitanio et al., 2003). For mixed variables, i.e. some variables are discrete and others are continuous, the conditional Gaussian distribution is used (Lauritzen and Wermuth, 1989).

Graphical models have been applied in a wide range of settings. One prominent application of graphical models is in probabilistic expert systems as described in Cowell et al. (1999). Other recent applications are Caputo et al. (1999, 2003), Corradi et al. (2003), Mohamed et al. (1998), Friedman et al. (2000), Waddell and Kishino (2000), Hartemink et al. (2001), Lauritzen and Sheehan (2003), and Wang et al. (2003). The edited volume Green et al. (2003) features further applications, as does a soon to appear review paper by Jordan (2004).

1.2 AMP chain graph models

Under the original LWF Markov property the chain graph depicted in Figure 1.1(c) specifies the conditional independences

$$1 \perp\!\!\!\perp 2, \quad 1 \perp\!\!\!\perp 4 \mid (2,3), \quad 2 \perp\!\!\!\perp 3 \mid (1,4), \tag{1.1}$$

where the labels $1, 2, 3, 4$ represent some random variables X_1, X_2, X_3, X_4. If X_1, X_2, X_3, X_4 are jointly multivariate normal, then the conditional independences $1 \perp\!\!\!\perp 4 \mid (2,3)$ and $2 \perp\!\!\!\perp 3 \mid (1,4)$ do not lead to constraints that are simply interpretable in terms of regression coefficients (Andersson et al., 2001, §1). However, Andersson et al. (2001) have introduced an alternative Markov property that, for multivariate normal variables, leads to simple constraints on regression coefficients.

For the chain graph in Figure 1.1(c) this AMP Markov property specifies

$$1 \perp\!\!\!\perp 2, \quad 1 \perp\!\!\!\perp 4 \mid 2, \quad 2 \perp\!\!\!\perp 3 \mid 1. \tag{1.2}$$

(These conditional independence are not specified by any undirected graph, any DAG, or any chain graph under the LWF Markov property.) Normal random variables X_1, X_2, X_3, X_4 satisfy the conditional independences (1.2) iff they fulfill the block-recursive regression equations (Andersson et al., 2001, Eqn. (3)):

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= \varepsilon_2 \\ X_3 &= \beta_1 X_1 + \varepsilon_3 \\ X_4 &= \beta_2 X_2 + \varepsilon_4, \end{aligned} \tag{1.3}$$

where β_1 and β_2 are scalar regression coefficients and ε_1, ε_2 and $(\varepsilon_3, \varepsilon_4)$ are independent normally distributed residuals with zero means, ε_1 and ε_2 have arbitrary variances and $(\varepsilon_3, \varepsilon_4)$ are distributed as bivariate normal with arbitrary positive definite covariance matrix.

Block-recursive regression equations similar to the ones in (1.3) can be associated with any general Gaussian AMP chain graph model (Andersson et al., 2001, §5), which immediately yields a parameterization of the models. Methodology for statistical inference in AMP chain graph models has not yet been developed but for Gaussian AMP chain graph models the block-recursive regression equations take on a generalized form of the well-known seemingly unrelated regressions = SUR (Zellner, 1962). The generalization consists in the fact that the residuals satisfy conditional independences specified by an undirected graph.

Two chapters of this thesis are designated to likelihood inference in Gaussian AMP chain graph models. In Chapter 2, we study in detail the likelihood equations for the Gaussian AMP chain graph model associated with the graph in Figure 1.1(c). Since for this simple model the conditional distribution $(X_3, X_4 \mid X_1, X_2)$ comprises all the complexity of likelihood inference, we formulate this chapter in terms of the conditional model, which falls precisely in the conventional framework of SUR. The results of Chapter 2 are partly negative since we show that the likelihood of this simple Gaussian AMP chain graph model may be multimodal. However, the results are also constructive because we are able to show that this model may not have more than five solutions to the likelihood equations, and that the probability of multimodality vanishes asymptotically for growing sample size.

In Chapter 3, we present an algorithm for solving the likelihood equations of an arbi-

trary Gaussian AMP chain graph model. Andersson et al. (2001, §5) state that maximum likelihood estimation in Gaussian AMP chain graph models requires "iterative methods for generalized SUR and covariance selection models [= Gaussian undirected graph models]" and in fact the algorithm we present in Chapter 3 is a combination of iterative proportional fitting, used to fit Gaussian undirected graph models, and generalized least squares, used to fit SUR models.

1.3 Ancestral graph models

Ancestral graphs, introduced by Richardson and Spirtes (2002), are generalizations of undirected graphs and DAGs that are different from chain graphs. Ancestral graphs are graphs with three types of edges: undirected and directed edges are complemented by bidirected edges. An example of an ancestral graph is shown in Figure 1.2

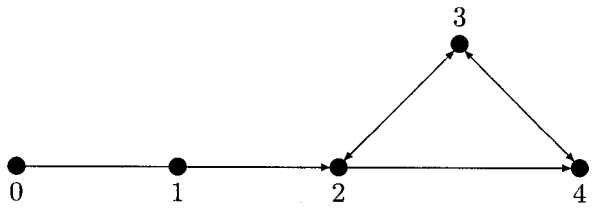


Figure 1.2: An ancestral graph.

In general, ancestral graphs specify conditional independences that are not specified by any AMP or LWF chain graph. This is the case, for example, for the graph shown in Figure 1.2, which states that

$$0 \perp\!\!\!\perp (2,3,4) \mid 1, \quad (0,1) \perp\!\!\!\perp 3, \quad (0,1) \perp\!\!\!\perp 4 \mid 2. \tag{1.4}$$

However, the conditional independences stated in (1.2), which are specified by the AMP chain graph in Figure 1.1(c), are also specified by the ancestral graph in Figure 1.3(a). This implies, in particular, that the Gaussian likelihood equations analyzed in Chapter 2 also stem from Gaussian ancestral graph models. Thus, the likelihood of Gaussian ancestral graph models may be multimodal.

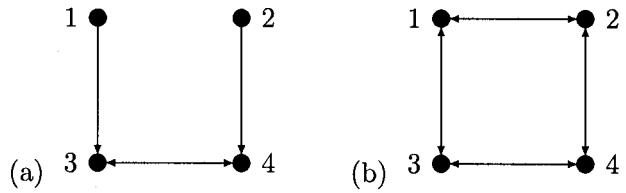


Figure 1.3: (a) An ancestral graph, (b) a bidirected graph.

Besides providing a generalization of undirected graphs and DAGs, ancestral graphs are designed to be able to encode all multivariate dependence structures that may arise from a DAG under conditioning and marginalization. Thus, the causal interpretation of DAGs (Pearl, 2000; Spirtes et al., 2000; Lauritzen, 2001) allows for a causal interpretation of ancestral graphs, see Richardson and Spirtes (2003).

In this thesis, we propose a new algorithm for maximum likelihood estimation in Gaussian ancestral graph models. In Chapter 4, we develop the new iterative conditional fitting algorithm for the special case of bidirected graphs, also termed covariance graphs (illustrated in Figure 1.3(b)). Subsequently, in Chapter 5, we generalize it to cover arbitrary ancestral graphs.

1.4 Summary of the chapter organization

This thesis considers maximum likelihood (ML) estimation of the parameters of AMP chain graph models and ancestral graph models in the case of continuous variables with a joint multivariate Gaussian distribution. In Chapter 2, we show that the likelihood of Gaussian AMP chain graph and ancestral graph models may be multimodal. In Chapter 3, we combine existing techniques (iterative proportional fitting, generalized least squares) into an algorithm for ML estimation in AMP chain graph models. A new algorithm for ML estimation in Gaussian ancestral graph models is presented in Chapters 4 and 5, where Chapter 4 deals with the special case of bidirected graphs and Chapter 5 treats general ancestral graphs. Appendix A provides general results on the convergence of the iterative algorithms proposed in this thesis.

Chapter 2

MULTIMODALITY OF THE LIKELIHOOD IN THE BIVARIATE SEEMINGLY UNRELATED REGRESSIONS MODEL

In this chapter, we analyze the simplest two-equation seemingly unrelated regressions model and demonstrate that its likelihood may have up to 5 stationary points, and thus there may be up to 3 local modes. Consequently the estimates obtained via iterative estimation methods may depend on starting values. We further show that the probability of multimodality vanishes asymptotically. Monte Carlo simulations suggest that multimodality rarely occurs if the seemingly unrelated regressions model is true, but can become more frequent if the model is misspecified. The existence of multimodality in the likelihood for seemingly unrelated regressions models contradicts several claims in the literature.

2.1 Introduction

In multivariate analysis of variance, correlated regression equations all involve the same covariates. Seemingly unrelated regressions are also correlated regression equations but allow different covariates in different equations. The seemingly unrelated regressions model 'plays a central role in contemporary econometrics' (Goldberger, 1991, p. 323) but also appears for longitudinal data (Rochon, 1996a,b; Verbyla and Venables, 1988). References to literature up to the mid-1980s can be found in Srivastava and Giles (1987). More recently, seemingly unrelated regressions came up in likelihood factorizations of Gaussian graphical models (Andersson et al., 2001; Richardson and Spirtes, 2002); compare Chapter 1. Bayesian approaches are reviewed, for example, by Percy (1996).

In his pioneering work, Zellner (1962, 1963) introduced a two-stage estimator and pointed out how the consideration of the correlation between regression equations leads to a gain in efficiency; see also Binkley and Nelson (1988). The two-stage estimator is a generalized least-

squares estimator for the regression coefficients based on the covariance matrix estimator found from residuals after an initial estimation of the regression coefficients. Iterating Zellner's two-stage procedure yields a sequence of estimators that converges to a solution to the Gaussian likelihood equations; for details see Oberhofer and Kmenta (1974), Meng and Rubin (1993, 1996) and §2.4 in this chapter.

This chapter studies multimodality of the likelihood in a bivariate Gaussian seemingly unrelated regressions model with two covariates because this two-equation case is simple enough to allow a thorough study of its nonlinear likelihood equations. In §2.2, we define the model and cite simulated data with a bimodal likelihood. As shown in §2.3, the model's likelihood equations almost surely have one, three or possibly five solutions corresponding to a unimodal, bimodal or trimodal likelihood. For increasing sample size, we prove that the probability of multimodality vanishes asymptotically. In §2.4, we comment on iterative maximum likelihood estimation. The Monte Carlo study in §2.5 investigates the frequency of multimodality. The discussion in §2.6 reviews and corrects claims in the literature.

2.2 The model and the simulated data

In bivariate seemingly unrelated regressions with two covariates, the only model for which maximum likelihood estimation is not straightforward is the model in which one response variable is regressed on the first covariate and the other response variable is regressed on the second covariate; compare Andersson and Perlman (1994). This model is the focus of this chapter and may be specified as follows.

Let $(X_{1j}, X_{2j}, Y_{1j}, Y_{2j})$, $j = 1, \ldots, n$, be independent and identically distributed random vectors with a joint normal distribution. We group the X_{ij} and the Y_{ij} in the $2 \times n$ random matrices X and Y, respectively. Furthermore, let X_i' and Y_i' be the ith rows of X and Y, respectively. Conditionally on the covariates X, the response variables Y are assumed to follow the seemingly unrelated regressions model

$$(Y \mid X) \sim N_{2\times n}\left\{\begin{pmatrix}\beta_1 & 0\\ 0 & \beta_2\end{pmatrix} X,\ \Sigma \otimes I_n\right\}. \tag{2.1}$$

Here Σ is the positive definite 2×2 conditional covariance matrix for the columns of Y

given X. Since X has a normal distribution in $\mathbb{R}^{2\times n}$ it is almost surely of full rank two whenever $n \geq 2$. Hence, we can assume X to be of full rank in the analysis of $(Y \mid X)$. The off-diagonal zeros in the mean parameter matrix of (2.1) imply that X_1 is conditionally independent of Y_2 given X_2 and X_2 is conditionally independent of Y_1 given X_1.

Sometimes it is convenient to rewrite the model (2.1) in terms of $\text{vec}(Y) = (Y_1', Y_2')$, the row-wise vectorization of Y to a row-vector of length $2n$. Let $\beta' = (\beta_1, \beta_2)$ and

$$Z = \begin{pmatrix} X_1' & 0 \\ 0 & X_2' \end{pmatrix} \in \mathbb{R}^{2\times 2n}.$$

Then (2.1) becomes

$$(\text{vec}(Y) \mid X) \sim N_{2n}\left(\beta' Z,\ \Sigma \otimes I_n\right). \tag{2.2}$$

Table 2.1 gives simulated data that, as we will show in §2.3, lead to a bimodal likelihood for model (2.1). In particular, these data demonstrate the possibility of a multimodal likelihood in Gaussian AMP chain graph models (Andersson et al., 2001) and Gaussian ancestral graph models (Richardson and Spirtes, 2002). In detail, the covariates were generated as

Table 2.1: Simulated data with a bimodal likelihood.

j	X_{1j}	X_{2j}	Y_{1j}	Y_{2j}
1	1.88	0.55	2.34	4.97
2	0.22	-2.16	-0.05	-3.26
3	-0.46	1.16	0.06	2.66
4	0.77	-0.30	1.82	-0.03
5	-1.03	1.31	-1.93	0.93
6	0.74	1.95	2.78	5.58
7	0.83	-3.11	0.62	-5.84
8	1.01	-2.39	-0.68	-2.24

$X \sim N_{2n}\{0,\ \text{diag}(1,5) \otimes I_n\}$ where $\text{diag}(1,5)$ is the 2×2 diagonal matrix with diagonal

entries 1 and 5, and the observations Y were simulated conditionally on X according to (2.1) with

$$\beta' = (1, 2), \quad \Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 2 \end{pmatrix}. \tag{2.3}$$

2.3 *Maximization of the likelihood*

2.3.1 The likelihood equations

The log-likelihood of (2.1), conditionally upon X, is

$$\ell(\beta, \Sigma) = -n \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \mathrm{tr} \left[\Sigma^{-1} \{Y - \mathrm{diag}(\beta) X\} \{Y - \mathrm{diag}(\beta) X\}' \right]. \tag{2.4}$$

If we use the form (2.2), the log-likelihood can be rewritten as

$$\ell(\beta, \Sigma) = -n \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \left\{ \mathrm{vec}(Y) - \beta' Z \right\} \left(\Sigma^{-1} \otimes I_n \right) \left\{ \mathrm{vec}(Y) - \beta' Z \right\}'. \tag{2.5}$$

Taking the derivatives with respect to β in (2.5) and the precision matrix Σ^{-1} in (2.4), we obtain the likelihood equations

$$\beta' = \mathrm{vec}(Y) \left(\Sigma^{-1} \otimes I_n \right) Z' \left\{ Z \left(\Sigma^{-1} \otimes I_n \right) Z' \right\}^{-1}, \tag{2.6}$$

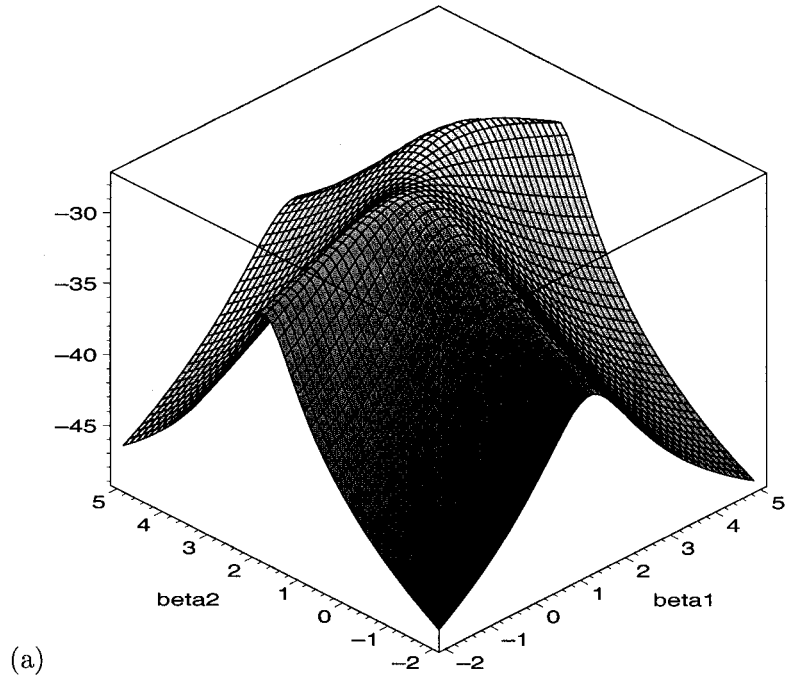
$$\Sigma = \frac{1}{n} \{Y - \mathrm{diag}(\beta) X\} \{Y - \mathrm{diag}(\beta) X\}'. \tag{2.7}$$

As is readily seen, the constrained log-likelihood $\ell_\Sigma : \beta \mapsto \ell(\beta, \Sigma)$ is strictly concave for any fixed positive definite Σ. Hence, (2.6) gives β which maximizes ℓ_Σ. Conversely, the constrained log-likelihood $\ell_\beta : \Sigma \mapsto \ell(\beta, \Sigma)$ is strictly concave in the precision matrix Σ^{-1}. Thus, (2.7) yields the maximizer of ℓ_β, which we call $\hat{\Sigma}(\beta)$. In particular, this implies that the log-likelihood $\ell(\beta, \Sigma)$ has no local minimum.

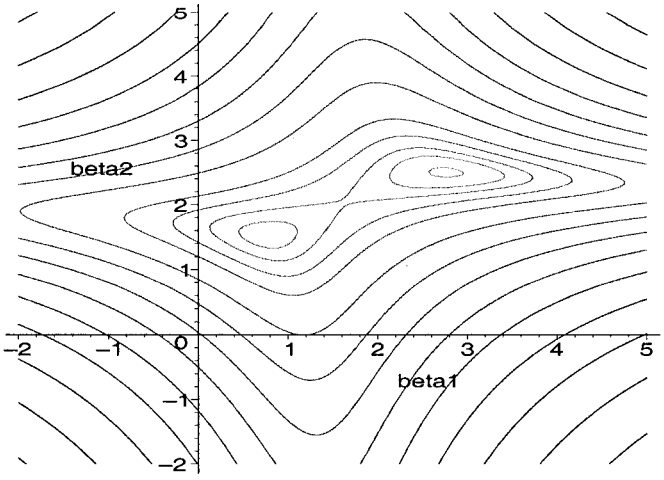
2.3.2 The profile likelihood

Since the parameter β is only two-dimensional we can plot the profile log-likelihood

$$\ell(\beta) = \max\{\ell(\beta, \Sigma) : \Sigma \text{ positive definite}\} = \ell\{\beta, \hat{\Sigma}(\beta)\}. \tag{2.8}$$



(a)



(b)

Figure 2.1: Data in Table 2.1. (a) Three-dimensional plot and (b) contour plot of the profile log-likelihood. The contour levels are at -48, -46, -44,..., -30, -29, -28.31, -27.9, and -27.53.

For the data in Table 2.1, the profile log-likelihood exhibits two local maxima, as can be seen in the three-dimensional plot and the contour plot in Figure 2.1.

The profile log-likelihood of a seemingly unrelated regressions model is of a simple form since, with β fixed, it is the maximized log-likelihood for multivariate normal observations with known mean. In more detail, plugging $\hat{\Sigma}(\beta)$ into the log-likelihood (2.4) yields

$$\ell(\beta) = -n\log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}(\beta)| - n. \tag{2.9}$$

Since the constrained log-likelihood ℓ_β is strictly concave for all β, the profile log-likelihood preserves all stationary points of the log-likelihood.

Lemma 2.1. *The parameter (β, Σ) is a stationary point of the log-likelihood if and only if β is a stationary point of the profile log-likelihood and $\Sigma = \hat{\Sigma}(\beta)$. Further, local maxima of the log-likelihood correspond to local maxima of the profile log-likelihood and vice versa.*

The maximum likelihood estimator $(\hat{\beta}, \hat{\Sigma})$ can be found by maximizing the profile log-likelihood to obtain $\hat{\beta}$ and using (2.7) to find $\hat{\Sigma} = \hat{\Sigma}(\hat{\beta})$. By (2.9), maximizing $\ell(\beta)$ is equivalent to minimizing the determinant $|\hat{\Sigma}(\beta)|$, which is a bivariate polynomial:

$$\begin{aligned}|\hat{\Sigma}(\beta)| \cdot n^2 = {} & \left(a_{22}\beta_2^2 - 2a_{21}\beta_2 + a_{20}\right)\beta_1^2 \\ & - \left(a_{12}\beta_2^2 - a_{11}\beta_2 + a_{10}\right)2\beta_1 + \left(a_{02}\beta_2^2 - 2a_{01}\beta_2 + a_{00}\right),\end{aligned}$$

where

$$\begin{aligned}
a_{22} &= \left|XX'\right| && , \quad a_{21} = \left|\begin{pmatrix} X_1' \\ Y_2' \end{pmatrix} X'\right| && , \quad a_{20} = \left|\begin{pmatrix} X_1' \\ Y_2' \end{pmatrix}\begin{pmatrix} X_1' \\ Y_2' \end{pmatrix}'\right|, \\
a_{12} &= \left|\begin{pmatrix} Y_1' \\ X_2' \end{pmatrix} X'\right| && , \quad a_{11} = \left|YX'\right| + \left|\begin{pmatrix} X_1' \\ Y_2' \end{pmatrix}\begin{pmatrix} Y_1' \\ X_2' \end{pmatrix}'\right|, \quad a_{10} = \left|Y\begin{pmatrix} X_1' \\ Y_2' \end{pmatrix}'\right|, \\
a_{02} &= \left|\begin{pmatrix} Y_1' \\ X_2' \end{pmatrix}\begin{pmatrix} Y_1' \\ X_2' \end{pmatrix}'\right|, \quad a_{01} = \left|Y\begin{pmatrix} Y_1' \\ X_2' \end{pmatrix}'\right| && , \quad a_{00} = \left|YY'\right|.
\end{aligned} \tag{2.10}$$

This polynomial is of fourth degree since $a_{22} > 0$ because X is of full rank. The coefficients a_{ij} do not range freely but are related through a number of inequalities that stem from the

fact that, if A and B are two $k \times n$ matrices, $k < n$, then $|AA'| \geq 0$ and $|AB'|^2 \leq |AA'||BB'|$. The latter inequality follows from the Cauchy-Binet determinant formula (Edwards, 1994, p. 326) and the Cauchy-Schwarz inequality. With probability one, the inequalities are strict, and thus, for example, a_{22}, a_{20}, a_{02} and a_{00} are positive, and $a_{21}^2 < a_{22}a_{20}$, $a_{12}^2 < a_{22}a_{02}$, $a_{10}^2 < a_{20}a_{00}$ and $a_{01}^2 < a_{02}a_{00}$.

The derivatives of $|\hat{\Sigma}(\beta)|$ with respect to β_1 and β_2 are linear in β_1 and β_2, respectively. Hence, we can determine the root of one derivative and plug it into the other derivative. The equation $\partial|\hat{\Sigma}(\beta)|/\partial\beta_2 = 0$ has the solution

$$\hat{\beta}_2(\beta_1) = \frac{a_{21}\beta_1^2 - a_{11}\beta_1 + a_{01}}{a_{22}\beta_1^2 - 2a_{12}\beta_1 + a_{02}}. \tag{2.11}$$

Plugging (2.11) into $\partial|\hat{\Sigma}(\beta)|/\partial\beta_1$ gives

$$\left.\frac{\partial|\hat{\Sigma}(\beta)|}{\partial\beta_1}\right|_{\beta=(\beta_1,\hat{\beta}_2(\beta_1))} = \frac{2}{n^2} \cdot \frac{f(\beta_1)}{\left(a_{22}\beta_1^2 - 2a_{12}\beta_1 + a_{02}\right)^2}.$$

Here f is a polynomial of degree five which does not seem to have an explicit factorization:

$$\begin{aligned} f(x) \;=\; & x^5 a_{22}\left(a_{20}a_{22} - a_{21}^2\right) \\ & + x^4\left(a_{22}a_{21}a_{11} - a_{10}a_{22}^2 - 4a_{22}a_{20}a_{12} + 3a_{21}^2 a_{12}\right) \\ & + x^3\left(4a_{22}a_{10}a_{12} + 2a_{22}a_{20}a_{02} - 4a_{21}a_{11}a_{12} - 2a_{21}^2 a_{02} + 4a_{20}a_{12}^2\right) \\ & + x^2\left(2a_{21}a_{01}a_{12} - a_{22}a_{01}a_{11} - 2a_{22}a_{10}a_{02} - 4a_{20}a_{12}a_{02} - 4a_{10}a_{12}^2\right. \\ & \qquad \left. + a_{11}^2 a_{12} + 3a_{21}a_{11}a_{02}\right) \\ & + x\left(a_{01}^2 a_{22} - a_{11}^2 a_{02} - 2a_{21}a_{01}a_{02} + a_{20}a_{02}^2 + 4a_{10}a_{12}a_{02}\right) \\ & + \left(a_{01}a_{11}a_{02} - a_{01}^2 a_{12} - a_{10}a_{02}^2\right). \end{aligned} \tag{2.12}$$

If β_1 is one of possibly five real roots of f then (2.11) gives a stationary point $\left(\beta_1, \hat{\beta}_2(\beta_1)\right)$ of the profile log-likelihood. The type of a stationary point is determined by the Hessian of $|\hat{\Sigma}(\beta)|$, which is the symmetric matrix

$$\frac{\partial^2|\hat{\Sigma}(\beta)|}{\partial\beta^2} = \frac{2}{n^2}\begin{pmatrix} \beta_2^2 a_{22} - 2\beta_2 a_{21} + a_{20} & 2\beta_1\beta_2 a_{22} - 2\beta_1 a_{21} - 2\beta_2 a_{12} + a_{11} \\ \star & \beta_1^2 a_{22} - 2\beta_1 a_{12} + a_{02} \end{pmatrix}. \tag{2.13}$$

The polynomial f for the data from Table 2.1 is plotted in Figure 2.2. It has three roots

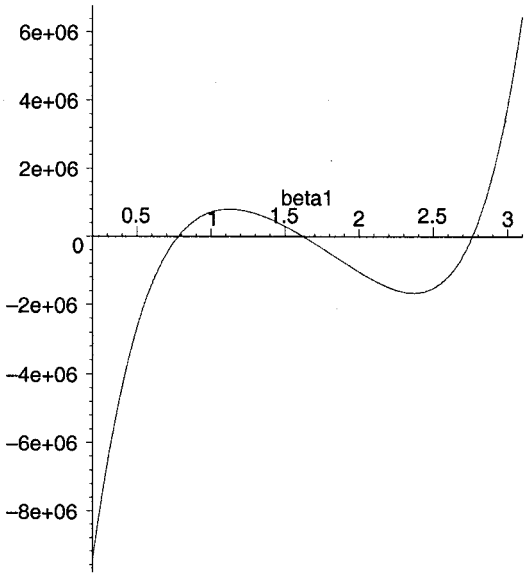


Figure 2.2: Plot of $f(\beta_1)$.

Table 2.2: Roots of f for data from Table 2.1.

j	$\beta_1(j)$	$\hat{\beta}_2(\beta_1(j))$	$\ell\{\beta(j)\}$
1	0.78	1.54	-27.73
2	1.62	2.03	-28.30
3	2.76	2.50	-27.49

$\beta_1(j)$, $j = 1, 2, 3$, which we cite in Table 2.2 with the corresponding values of $\hat{\beta}_2(\beta_1)$ and the profile log-likelihood. The Hessian (2.13) is positive definite for $\beta_1(1)$ and $\beta_1(3)$ and indefinite for $\beta_1(2)$. Hence, for $j = 1, 3$, we get the two local maxima and $j = 2$ gives us the saddle point of the log-likelihood. Since, almost surely, the entry (2,2) in the Hessian (2.13) is positive for all β_1 we maximize $\ell\{(\beta_1, \beta_2)\}$ for a given β_1 by setting $\beta_2 = \hat{\beta}_2(\beta_1)$, i.e.

$$\ell(\beta_1) = \ell\{(\beta_1, \hat{\beta}_2(\beta_1))\} = \max[\ell\{(\beta_1, \beta_2)\} : \beta_2 \in \mathbb{R}]. \tag{2.14}$$

An analog to Lemma 2.1 applies also to this 'profile profile' log-likelihood $\ell(\beta_1)$, which as a smooth univariate function has a local minimum between two local maxima. Furthermore,

local minima of $\ell(\beta_1)$ correspond to saddle points of the likelihood. This establishes the following result.

Theorem 2.2. *The two-equation seemingly unrelated regressions model (2.1) almost surely has one, three or five solutions to the likelihood equations. They are determined by the real roots of the fifth degree polynomial f defined in (2.12). Given a real root β_1 of f the solution to the likelihood equations can be completed by setting $\beta_2 = \hat{\beta}_2(\beta_1)$ from (2.11) and $\Sigma = \hat{\Sigma}(\beta)$ from (2.7). If $\{\beta_1(j)\}_j$ are the ordered real roots of f then the roots for odd j lead to local maxima of the likelihood and the roots for even j to saddle points.*

Remark 1. Sugiura and Gupta (1987) prove a similar result involving a third degree polynomial in the Behrens-Fisher problem of estimating a common mean of two normal populations with unequal unknown variances.

2.3.3 Large sample size asymptotics

The following theorem asserts that the number of solutions to the likelihood equations converges almost surely to one. This suggests that multimodality of the likelihood predominantly occurs with small sample sizes, when the model is correct.

Theorem 2.3. *If $X \sim N_{2n}(0, \Sigma^X \otimes I_n)$ has a centered normal distribution and $(Y \mid X)$ follows the two-equation seemingly unrelated regressions model (2.1) then the number of solutions to the likelihood equations of (2.1) converges almost surely to one.*

Proof. To avoid notational confusion, let $\mathrm{var}(Y \mid X) = \Sigma^Y \otimes I_n$ in (2.1). Furthermore, let $\Lambda \otimes I_n = \mathrm{var}(X, Y)$ be the joint covariance matrix which can be easily reconstructed from Σ^X, Σ^Y and β. The law of large numbers applied to the coefficients a_{ij} from (2.10) yields that, as n goes to infinity, a_{ij}/n^2 converges almost surely to the determinant of the submatrix of Λ which corresponds to a_{ij}. Thus, for example, $a_{22}/n^2 \to |\Lambda_{(12),(12)}| = |\Sigma^X|$ almost surely. If we plug all the almost sure limits for the a_{ij} into the polynomial f from (2.12) then we obtain that $f(x)/n^2$ converges almost surely to

$$f^{\mathrm{asy}}(x) = (x - \beta_1)\left[\sigma^X_{11}\sigma^Y_{22}\left\{|\Sigma^X|(x-\beta_1)^2 + \sigma^Y_{11}\sigma^X_{22}\right\}^2 - \sigma^Y_{11}\sigma^X_{22}\left(\sigma^X_{12}\right)^2\left(\sigma^Y_{12}\right)^2\right]. \quad (2.15)$$

Obviously, $f^{\text{asy}}(x)$ has the real root β_1. The other four roots are complex because $|\Sigma^X| > 0$ and $|\Sigma^Y| > 0$. Since the locations of the complex roots of a polynomial depend continuously on its coefficients, the number of real roots of $f(x)$ converges to one almost surely. □

Remark 2. This result seems at first sight obvious but it does not follow from general theory. For a related result in the Cauchy location family, see Reeds (1985).

2.4 Iterative maximum likelihood estimation

The maximum likelihood estimator $(\hat{\beta}, \hat{\Sigma})$ can be found by iterating Zellner's two-stage procedure, i.e. by iteratively solving (2.6) and (2.7) for fixed Σ and β, respectively. By definition, a fixed point of the resulting algorithm solves the likelihood equations, and, since (2.6) and (2.7) correspond to maximizing a constrained log-likelihood, the iteration steps never decrease the log-likelihood. Moreover, if the sample size n is larger than or equal to the number of response variables plus the number of covariates, here if $n \geq 4$, then with probability one the observations are such that the set

$$\Theta = \{(\beta, \Sigma) \mid \beta \in \mathbb{R}^2,\ \Sigma \in \mathbb{R}^{2\times 2} \text{ positive definite},\ \ell(\beta, \Sigma) \geq \ell(\beta^{(0)}, \Sigma^{(0)})\}$$

is compact. Here $\beta^{(0)}$ and $\Sigma^{(0)}$ are arbitrary starting values. Therefore, the sequence of estimators remains in the compact set Θ and all its accumulation points are solutions to the likelihood equations that all have the same likelihood. In fact, Corollary A.5 in Appendix A yields that the sequence of estimators converges because we have shown that the likelihood equations for the model (2.1) may have only finitely many solutions. This convergence property appears contrary to a remark by Cox and Wermuth (1996, p. 130) that there is no guarantee of convergence.

Two starting values are traditionally used. The identity matrix as initial estimate of Σ results in the 'restricted' iterative seemingly unrelated regressions estimate ISURR; the estimate of Σ from the multivariate analysis of variance model

$$(Y \mid X) \sim N_{2n}\left\{ \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} X,\ \Sigma \otimes I_n \right\} \tag{2.16}$$

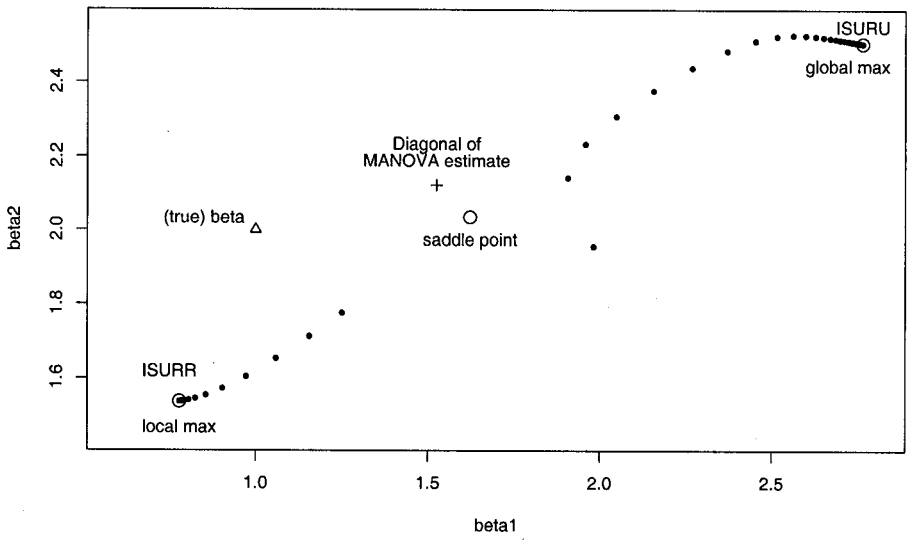


Figure 2.3: The sequences of estimates generated by ISURR and ISURU.

yields the 'unrestricted' estimate ISURU. The name 'unrestricted' reflects that β_{12} and β_{21} are not restricted to zero; the term 'restricted' contrasts this. Note that the first estimate of β in the restricted approach is the estimate obtained by applying ordinary least squares separately to the two regressions $(Y_1 \mid X_1)$ and $(Y_2 \mid X_2)$.

For the data in Table 2.1, the sequences of restricted and unrestricted estimates converge to the local maxima based on $\beta_1(1)$ and $\beta_1(3)$ from Table 2.2, respectively. Figure 2.3 illustrates the sequences generated by ISURR and ISURU. The covariance matrices are estimated to be

$$\hat{\Sigma}_{\text{ISURR}} = \begin{pmatrix} 1.35 & 1.17 \\ 1.17 & 3.62 \end{pmatrix}, \qquad \hat{\Sigma}_{\text{ISURU}} = \begin{pmatrix} 3.34 & -3.77 \\ -3.77 & 5.24 \end{pmatrix}.$$

Hence, the mean parameters and the variances are estimated to quite different values and the estimated covariances even have different signs. The global maximum is found by the unrestricted method but when compared to the true parameters (2.3) the estimate from the restricted method is preferable.

Theoretically, the algorithm could also converge to the saddle point; for example starting in the saddle point it would never leave it. However, in practice the finite accuracy in computer calculations seems to prevent convergence to a saddle point.

2.5 Monte Carlo study

Theorem 2.2 permits us to investigate the finite-sample probability of multimodality in Monte Carlo simulations. We repeatedly simulate covariates $X \sim N_{2n}\{0,\ \mathrm{diag}(1,5) \otimes I_n\}$ and response variables Y conditionally upon X according to (2.1) with the same parameters as in (2.3) with the exception of covariance σ_{12} which we vary to take on the values $\{0, \pm 0.6, \pm 1.3\}$ corresponding to the correlation coefficients $\rho \in \{0, \pm 0.42, \pm 0.92\}$. Table 2.3 summarizes the results of 10,000 simulations, which show that multimodality is relatively rare and predominantly occurs with small sample sizes, as Theorem 2.3 suggested. For example, for $n = 8$, the chance of bimodality is about 1 in 1000. Trimodality occurred only once for $n = 5$ and $\rho = 0.92$. The iterative restricted and unrestricted estimators converge to different modes in roughly one third of the multimodal cases.

Sometimes, however, the data might not come from seemingly unrelated regressions but from a misspecified model. It could for example be the case that the true model is the MANOVA model (2.16) which is a super-model of the seemingly unrelated regressions model (2.1). We simulate the covariates again as $X \sim N_{2n}\,(0,\ \mathrm{diag}(1,5) \otimes I_n)$ but now the observations Y conditionally upon the covariates are simulated as $N_{2n}\,(BX,\ \Sigma \otimes I_n)$ where B is a 2×2 mean parameter matrix of our choice. We choose the values

$$B = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.4 \\ 1.8 & 2 \end{pmatrix},$$

which are motivated by the maximum likelihood estimates in the MANOVA model (2.16) for the observations in Table 2.1. Further, we fix the (conditional) variances $\sigma_{11} = 1$, and $\sigma_{22} = 2$, but again vary the covariance σ_{12} to take on the values $0, \pm 0.6, \pm 1.3$.

Table 2.4 summarizes the results of 10,000 simulations. The chance of encountering multimodality in this mis-specification setting is of the order 1 in 10. Trimodality remains rare with only 3 observed examples. Since trimodality is so rare, we cite the observations

Table 2.3: Frequencies N_b of bimodality and N_i of the occurrence of unequal restricted and unrestricted estimates in 10,000 simulations from the seemingly unrelated regressions model (2.1). One trimodal likelihood function for $n = 5$ and $\rho = 0.92$, indicated by '+1'.

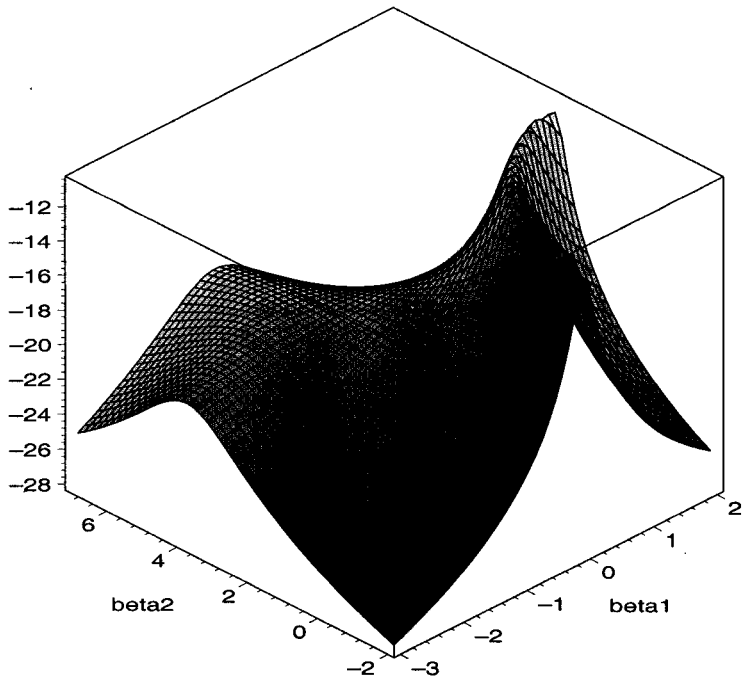
	$\rho = -0.92$		$\rho = -0.42$		$\rho = 0$		$\rho = 0.42$		$\rho = 0.92$	
n	N_b	N_i	N_b	N_i	N_b	N_i	N_b	N_i	N_b	N_i
5	261	80	297	127	328	166	299	131	262+1	83
6	79	21	97	49	120	53	100	45	96	25
7	20	7	38	17	44	19	46	21	29	13
8	16	5	14	6	23	15	12	3	7	2
9	3	2	8	5	11	3	3	1	2	0
10	1	0	4	0	2	1	6	4	0	0
11	1	1	0	0	1	1	2	0	0	0
12	0	0	0	0	0	0	1	1	0	0
13	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0

we found for $n = 5$ and $\rho = 0.92$ in Table 2.5. Figure 2.4 shows a three-dimensional plot and a contour plot of the profile log-likelihood for these observations.

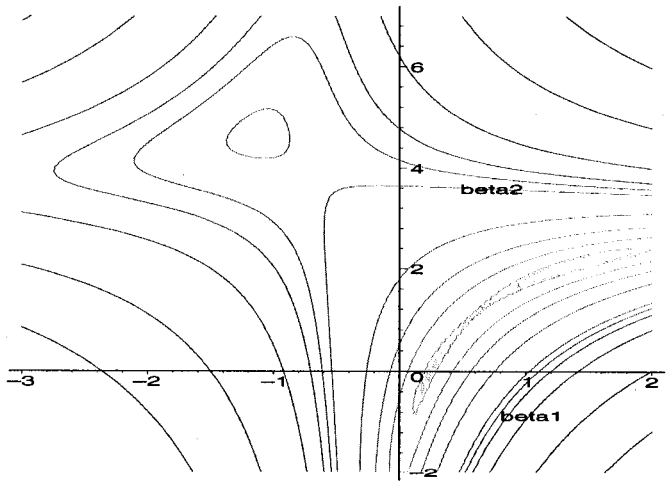
2.6 Related work and discussion

In §§2.2–2.5, we showed that the likelihood of a simple two-equation seemingly unrelated regressions model can be multimodal. We believe this holds true also for more complex seemingly unrelated regressions models since the profile log-likelihood will still be a monotone transformation of a multivariate polynomial in the mean parameters. It is surprising that the literature on seemingly unrelated regressions makes no clear statement about the possibility of multimodality.

In a recent textbook, for example, Greene (1997, p. 686) states incorrectly that 'the



(a)



(b)

Figure 2.4: Data in Table 2.5. (a) Three-dimensional plot and (b) contour plot of the profile log-likelihood. The contour levels are at -25, -23, -21, -20, -19.5, -19.05, -17, -15, -13, -11.2, -11, and -10.7.

Table 2.4: Frequencies N_b of bimodality and N_i of the occurrence of unequal restricted and unrestricted estimates in 10,000 simulations from the MANOVA (2.16). Three trimodal likelihood functions indicated as "+1" and "+2".

	$\rho = -0.92$		$\rho = -0.42$		$\rho = 0$		$\rho = 0.42$		$\rho = 0.92$	
n	N_b	N_i	N_b	N_i	N_b	N_i	N_b	N_i	N_b	N_i
5	2803	1363	3228+1	1231	3362	1233	3286	1375	2698	1335
6	2095+2	1032	2957	1022	2991	1016	2869	1044	2172	1056
7	1721	830	2563	839	2771	852	2660	917	1729	867
8	1448	717	2421	813	2634	731	2409	818	1459	679
9	1253	563	2220	723	2617	722	2167	725	1174	592
10	994	446	2217	732	2467	683	2145	703	985	423
11	831	382	2033	664	2390	621	2045	662	874	389
12	691	309	1949	623	2338	590	1915	620	734	348
13	599	255	1909	638	2208	558	1879	618	611	265
14	544	224	1717	520	2230	549	1741	578	509	208
15	402	153	1659	532	2244	545	1592	504	430	157

log-likelihood [of a seemingly unrelated regressions model] is globally concave' which is contradicted by our results. The claim would hold in a regular exponential family but, since seemingly unrelated regressions are defined by linear restrictions on the mean parameters and not on the natural parameters of the multivariate normal distribution, it forms only a curved exponential family; compare van Garderen (1997). This claim does not appear in the 2003 edition of the textbook but Greene does not go on to make any explicit statement about the possibility of a multimodal likelihood.

In §2.4, we showed that, for the simulated data from Table 2.1, the iterative version of Zellner's two-stage estimator converges to the two different local maxima depending on which one of the two usual starting values is used. Both of these starting values are consistent, if the starting value for Σ in the restricted approach is taken to be the estimate of Σ after one iteration. The fact that starting from two consistent estimates the algorithm yields

Table 2.5: Simulated data with a trimodal likelihood.

j	X_{1j}	X_{2j}	Y_{1j}	Y_{2j}
1	-0.65	-0.04	0.14	0.52
2	-0.80	-1.18	-0.73	-1.93
3	1.34	1.98	1.40	3.02
4	-1.03	-2.42	-2.29	-6.67
5	-1.08	-3.75	-3.30	-9.94

two different estimates contradicts a claim of Srivastava and Giles (1987, p. 157) who state that the algorithm's limit is unique if the initial estimate is consistent. Similarly, this fact contradicts Magnus (1978, Lemma 2), who states that the root of the likelihood equations produced by the iterations is the unique maximum likelihood estimator. The maximum likelihood estimator is trivially almost surely unique when defined by the global maximum of the likelihood. Solutions to the likelihood equations, however, are not necessarily unique. To justify their claims, Magnus (1978, Lemma 2) and Srivastava and Giles (1987, p. 157) both refer to general maximum likelihood estimation theory, the former to Cramér (1999) and Dhrymes (1970, §3), and the latter to Rao (1973, §5), that obviously cannot provide the required uniqueness.

Our study suggests that, in order to explore the likelihood as fully as possible, one should always try different starting values for iterative procedures. Note that Wu and Perlman (2000) find improved starting values, compare also Andersson and Perlman (1994), though in our simple two-equation model (2.1) their idea reduces to the restricted approach. Detecting the multimodality is especially important in likelihood ratio testing since finding a local, but not global, maximum of the likelihood would favor the larger model that is hypothesized against the seemingly unrelated regressions model. For model search, our results suggest first searching among a class of models in which multimodality does not arise in order to avoid stopping the procedure erroneously because of a seemingly poor fit of a seemingly unrelated regressions model caused by multimodality. Note also that

in Bayesian inference of seemingly unrelated regressions the posterior distribution might inherit multimodality from the likelihood.

In our Monte Carlo study multimodality occurred only rarely even for small sample sizes. However, this may not be true in more complex situations. Moreover, we observed that data from misspecified models can lead to a multimodal likelihood at higher frequencies; compare also Cox and Wermuth (1996, pp. 102-3). We conducted further Monte Carlo experiments in the correctly-specified model to determine whether or not one of the solutions to the likelihood equations is preferable in terms of mean square error, but no clear pattern emerged: the global maximum was preferable for a strong correlation of the two seemingly unrelated regression equations but for weak or no correlation the saddle point in between the local and the global maximum was preferable.

Chapter 3

FITTING GAUSSIAN AMP CHAIN GRAPH MODELS

The AMP Markov property is a recently proposed alternative Markov property for chain graphs (Andersson et al., 2001). In the case of continuous variables with a joint multivariate Gaussian = normal distribution, it is the AMP rather than the earlier introduced LWF Markov property that is coherent with data-generation by natural block-recursive regressions. This chapter considers Gaussian AMP chain graph models, for which we show that maximum likelihood estimation can be performed by combining generalized least squares and iterative proportional fitting to an iterative algorithm.

3.1 Introduction

Let $G = (V, E)$ be a graph with vertex set V and edge set E, where E comprises directed ($\rightarrow$) and/or undirected edges ($-$) such that there is at most one edge between any pair of vertices $i, j \in V$. Such a graph is called a *chain graph* if it has no partially directed cycles, which are paths from one vertex to itself in which there is at least one directed edge and all directed edges point in the same direction. A chain graph is shown in Figure 3.1.

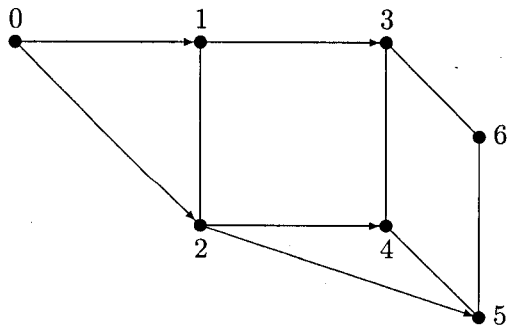


Figure 3.1: A chain graph with chain components $\{0\}$, $\{1, 2\}$ and $\{3, 4, 5, 6\}$.

Chain graphs have been introduced to provide conditional independence models that are more general than the ones obtained from undirected graphs or DAGs. The original LWF Markov property for chain graphs is due to Lauritzen and Wermuth (1989), Wermuth and Lauritzen (1990) and Frydenberg (1990). The parameterization of Gaussian LWF chain graph models is described in Wermuth (1992) and Lauritzen (1996, §5.4.1), and the maximum likelihood estimator (MLE) can be obtained by using the iterative proportional fitting algorithm (Speed and Kiiveri, 1986; Whittaker, 1990, pp.182–185). More precisely, iterative proportional fitting is applied multiple times to undirected graphs over overlapping subsets of the vertex set; compare Lauritzen (1996, §6.5.2).

The LWF chain graph approach is discussed, for example, in Didelez et al. (2002), Edwards (2000, §7.2) and Lauritzen and Richardson (2002). Examples of applications of LWF chain graph modelling are Mohamed et al. (1998) and Caputo et al. (1999, 2003).

Recently, Andersson et al. (2001) have proposed an alternative Markov property for chain graphs. In the case of continuous variables with a joint multivariate Gaussian = normal distribution, it is their AMP rather than the earlier introduced LWF Markov property that is coherent with data-generation by natural block-recursive regressions (Andersson et al., 2001, §§1 and 5). However, methodology for statistical inference in AMP chain graph models has not yet been developed. As a first step, this chapter provides an algorithm for ML estimation in Gaussian AMP chain graph models.

The chapter uses the notation of Andersson et al. (2001) and is organized as follows. We review the AMP Markov property and the parameterization of the associated Gaussian models in §§3.2 and 3.3, respectively. In §3.4, we derive the likelihood equations, which immediately suggest an iterative algorithm for ML estimation as described in §3.5. This algorithm combines generalized least squares and iterative proportional fitting. We conclude in §3.6.

3.2 The AMP Markov property for chain graphs

For a chain graph $G = (V, E)$, let $G^{\wedge}$ be the undirected graph obtained by deleting all directed edges in G. Let $\mathcal{T}$ be the set of connected components of $G^{\wedge}$. An element $\tau \in \mathcal{T}$ is

called a *chain component* of G. Clearly, the chain components form a partition of the vertex set $V = \dot{\cup}(\tau \mid \tau \in \mathcal{T})$. For the graph in Figure 3.1, for example, the chain components are $\tau_1 = \{0\}$, $\tau_2 = \{1, 2\}$ and $\tau_3 = \{3, 4, 5, 6\}$.

Associated with a chain graph G is an acyclic directed graph (ADG = DAG) over the chain components, i.e., a DAG $\mathcal{D} = \mathcal{D}(G)$ with vertex set $\mathcal{T}$. The DAG $\mathcal{D}$ comprises a directed edge $\tau \to \tau'$ if there exist vertices i and j in V such that $i \in \tau$, $j \in \tau'$, and $i \to j$ in G. The DAG $\mathcal{D}$ for the chain graph G from Figure 3.1 is shown in Figure 3.2.

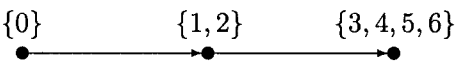


Figure 3.2: The DAG $\mathcal{D}(G)$ for G from Figure 3.1.

For a vertex $i \in V$, let $\mathrm{pa}_G(i) = \{j \in V \mid j \to i \in G\}$ be the *parents* of i in G. For a chain component $\tau \in \mathcal{T}$, let $\mathrm{pa}_{\mathcal{D}}(\tau) = \cup(\mathrm{pa}_G(i) \mid i \in \tau)$ be the parents of τ in $\mathcal{D}$. We write $\tau < \tau'$ if there exists a directed path $\tau \to \cdots \to \tau'$ in $\mathcal{D}$; otherwise we write $\tau \not< \tau'$. The *non-descendants* of τ in $\mathcal{D}$ are defined as $\mathrm{nd}_{\mathcal{D}}(\tau) = \{\tau' \mid \tau \not< \tau'\}$.

Now we can state the AMP Markov property in its block-recursive form.

Definition 3.1 (Andersson et al., 2001, Def. 5). *The AMP block-recursive Markov property specifies:*

C1: $\forall \tau \in \mathcal{T}: \quad \tau \perp\!\!\!\perp \big(\mathrm{nd}_{\mathcal{D}}(\tau) \setminus \mathrm{pa}_{\mathcal{D}}(\tau)\big) \,|\, \mathrm{pa}_{\mathcal{D}}(\tau)$;

C2: $\forall \tau \in \mathcal{T}: \quad$ *the conditional distribution* $\big(\tau \mid \mathrm{pa}_{\mathcal{D}}(\tau)\big)$ *is globally Markov with respect to the undirected subgraph* G_τ;

C3: $\forall \tau \in \mathcal{T}, \forall \sigma \subseteq \tau: \quad \sigma \perp\!\!\!\perp \big(\mathrm{pa}_{\mathcal{D}}(\tau) \setminus \mathrm{pa}_G(\sigma)\big) \,|\, \mathrm{pa}_G(\sigma)$.

The conditional independences in Definition 3.1 are stated in terms of sets of vertices, but in the sequel the vertices will be identified with random variables translating the statements in terms of vertices into proper conditional independence statements. For the graph G shown in Figure 3.2, the AMP block-recursive Markov property states, for example:

$(3,4,5,6) \perp\!\!\!\perp 0 \mid (1,2)$ from C1;

$3 \perp\!\!\!\perp 5 \mid (1,2,4,6)$ and $4 \perp\!\!\!\perp 6 \mid (1,2,3,5)$ from C2;

$3 \perp\!\!\!\perp 2 \mid 1$, $(4,5) \perp\!\!\!\perp 1 \mid 2$ and $6 \perp\!\!\!\perp (1,2)$ from C3.

A pairwise and a global version of the AMP Markov property are defined in Andersson et al. (2001) and Levitz et al. (2001, 2003), but in the Gaussian case considered subsequently these are all equivalent to the block-recursive Markov property stated in Definition 3.1. Here, two Markov properties are equivalent if for any graph the two sets of normal distributions satisfying the conditional independences stated by the respective Markov property are equal.

3.3 Gaussian AMP chain graph models

Let $X = (X_i \mid i \in V) \in \mathbb{R}^V$ be a random vector distributed according to the multivariate normal distribution $\mathcal{N}_V(0, \Sigma)$ where Σ is the unknown positive definite $V \times V$ covariance matrix. If we identify vertex sets $A \subseteq V$ with the random vector $X_A = (X_i \mid i \in A)$, then the statements of the AMP block-recursive Markov property become meaningful conditional independence statements imposing restrictions on Σ.

For a chain graph $G = (V, E)$, let $\mathbf{P}(G)$ be the set of all positive definite covariance matrices Σ that are such that $X \sim \mathcal{N}_V(0, \Sigma)$ satisfies all the conditional independences stated by the AMP block-recursive Markov property. The *Gaussian AMP chain graph model* is the family of normal distributions

$$\mathbf{N}(G) = \big(\mathcal{N}_V(0, \Sigma) \mid \Sigma \in \mathbf{P}(G)\big). \tag{3.1}$$

Now assume that we observe an i.i.d. sample $X^{(k)}$, $k \in N = \{1, \ldots, n\}$, from the Gaussian AMP chain graph model $\mathbf{N}(G)$ defined in (3.1). Let the vectors $X^{(k)}$ in the sample be the columns of the $V \times N$ random matrix Y, which then is distributed as

$$Y \in \mathbb{R}^{V \times N} \sim \mathcal{N}_{V \times N}(0, \Sigma \otimes I_N). \tag{3.2}$$

Here, I_N is the $N \times N$ identity matrix and $\otimes$ is the Kronecker product. The i-th row $Y_i \in \mathbb{R}^N$ of the matrix Y contains the observations for variable $i \in V$, made i.i.d. on all the subjects in N. The sample size is $n = |N|$ and the number of variables is $p = |V|$.

Since our model assumes a zero mean, the empirical covariance matrix, which is a sufficient statistic, is defined to be

$$S = \frac{1}{n} YY' \in \mathbb{R}^{V \times V}. \tag{3.3}$$

We shall assume that $n \geq p$ such that S is positive definite with probability one. The case where the model also includes an unknown mean vector $\mu \in \mathbb{R}^V$ can be treated by estimating μ by the empirical mean vector $\bar{Y} \in \mathbb{R}^V$, i.e. the vector of row means of Y. The empirical covariance matrix would then be replaced by the matrix

$$\tilde{S} = \frac{1}{n}(Y - \bar{Y} \otimes 1_N)(Y - \bar{Y} \otimes 1_N)' \in \mathbb{R}^{V \times V}, \tag{3.4}$$

where $1_N = (1, \ldots, 1) \in \mathbb{R}^N$. However, in this case only $n \geq p + 1$ would ensure that $\tilde{S}$ is positive definite with probability one.

The definition of the model $\mathbf{N}(G)$ in (3.1) is implicit. An explicit parameterization of $\mathbf{N}(G)$ is described in Andersson et al. (2001, §5). This parameterization associates one parameter with each vertex in V and each edge in E. Let $\Lambda = (\lambda_{ij})$ be a positive definite $V \times V$ matrix such that $\lambda_{ij} \neq 0$ only if $i = j$ or $i - j$. Let $B = (\beta_{ij})$ be an arbitrary real $V \times V$ matrix such that $\beta_{ij} \neq 0$ only if $j \to i$. With the parameter matrices B and Λ, we can define the covariance matrix

$$\Sigma = \Sigma(B, \Lambda) = (I_V - B)^{-1} \Lambda^{-1} ((I_V - B)^{-1})', \tag{3.5}$$

which is obviously positive definite. The matrices B and Λ parameterize the model $\mathbf{N}(G)$ since $\mathcal{N}_V(0, \Sigma) \in \mathbf{N}(G)$ iff there exist B and Λ such that $\Sigma = \Sigma(B, \Lambda)$, compare (3.11) and Andersson et al. (2001, eqn. (25)).

The directed graph $\mathcal{D}$ of chain components, defined in §3.2, is acyclic and, by C1 in Definition 3.1, the density of $X \sim \mathcal{N}_V(0, \Sigma) \in \mathbf{N}(G)$ can be factored into a product of conditional densities as

$$f(x) = \prod \left(f(x_\tau \mid x_{\mathrm{pa}_\mathcal{D}(\tau)}) \mid \tau \in \mathcal{T} \right), \quad x \in \mathbb{R}^V, \tag{3.6}$$

compare Andersson et al. (2001, eqn. (20)). The conditional distribution of $(X_\tau \mid X_{\mathrm{pa}_\mathcal{D}(\tau)})$ is

$$(X_\tau \mid X_{\mathrm{pa}_\mathcal{D}(\tau)}) \sim \mathcal{N}(\beta_\tau X_{\mathrm{pa}_\mathcal{D}(\tau)}, \Lambda_\tau), \quad \tau \in \mathcal{T}, \tag{3.7}$$

which is a "block-regression" for the "block" of variables τ. In (3.7), β_τ is the $\tau \times \mathrm{pa}_\mathcal{D}(\tau)$ submatrix of B and

$$\Lambda_\tau = \Sigma_{\tau,\tau} - \Sigma_{\tau,\mathrm{pa}_\mathcal{D}(\tau)}(\Sigma_{\mathrm{pa}_\mathcal{D}(\tau),\mathrm{pa}_\mathcal{D}(\tau)})^{-1}\Sigma_{\mathrm{pa}_\mathcal{D}(\tau),\tau} \tag{3.8}$$

is the $\tau \times \tau$ conditional covariance matrix. Note that Λ_τ is indeed equal to the $\tau \times \tau$ submatrix of Λ. The family of conditional parameters $((\Lambda_\tau, \beta_\tau) \mid \tau \in \mathcal{T})$ are called the family of G-parameters of $\Sigma \in \mathbf{P}(G)$.

The parameter space for the G-parameter β_τ is the set $\mathbf{B}_\tau(G)$ of all matrices β_τ that satisfy the condition

$$i \in \tau, j \in \mathrm{pa}_\mathcal{D}(\tau) \setminus \mathrm{pa}_G(i) \implies (\beta_\tau)_{ij} = 0. \tag{3.9}$$

Similarly the parameter space for the G-parameter Λ_τ is the set $\mathbf{P}(G_\tau)$ of all positive definite matrices Λ_τ that satisfy the condition

$$i, j \in \tau, i \not\sim j \text{ in } G \implies \lambda_\tau^{ij} = (\Lambda_\tau^{-1})_{ij} = 0. \tag{3.10}$$

As stated in Andersson et al. (2001, eqn. (25)), the parameter space of $\mathbf{N}(G)$ factors into a Cartesian product of parameter spaces for the G-parameters according to the bijection

$$\begin{aligned} \mathbf{P}(G) &\to \times(\mathbf{P}(G_\tau) \times \mathbf{B}_\tau(G) \mid \tau \in \mathcal{T}) \\ \Sigma &\mapsto ((\Lambda_\tau, \beta_\tau) \mid \tau \in \mathcal{T}). \end{aligned} \tag{3.11}$$

This parameter space factorization in conjunction with the density factorization (3.6) yields that the MLE of (B, Λ) can be obtained by separately computing the MLE for the parameters β_τ and Λ_τ in the block-regression (3.7).

3.4 Likelihood equations for a block-regression

The log-likelihood function $\ell(\beta_\tau, \Lambda_\tau)$ for the block-regression model from (3.7) can be expressed as

$$\ell(\beta_\tau, \Lambda_\tau) = -\frac{n|\tau|}{2}\log(2\pi) + \frac{n}{2}\log|\Lambda_\tau^{-1}| - \frac{n}{2}\mathrm{tr}\{\Lambda_\tau^{-1}S(\beta_\tau)\}, \tag{3.12}$$

where

$$S(\beta_\tau) = S_{\tau,\tau} - \beta_\tau S_{\mathrm{pa}_\mathcal{D}(\tau),\tau} - S_{\tau,\mathrm{pa}_\mathcal{D}(\tau)}\beta_\tau' + \beta_\tau S_{\mathrm{pa}_\mathcal{D}(\tau),\mathrm{pa}_\mathcal{D}(\tau)}\beta_\tau'. \tag{3.13}$$

Let $i, j \in \tau$ such that $i - j$ in G. It follows immediately from (3.12) that

$$\frac{\partial \ell(\beta_\tau, \Lambda_\tau)}{\partial \lambda_\tau^{ij}} = \frac{n}{2}\left((\Lambda_\tau)_{ij} - S(\beta_\tau)_{ij}\right), \tag{3.14}$$

where λ_τ^{ij}, $(\Lambda_\tau)_{ij}$, and $S(\beta_\tau)_{ij}$ are the ij-th entries in Λ_τ^{-1}, Λ_τ, and $S(\beta_\tau)$, respectively.

Let

$$d_\tau = |\{(i,j) \mid i \in \tau, j \in \mathrm{pa}_\mathcal{D}(\tau), j \to i\}| \tag{3.15}$$

be the number of directed edges between $\mathrm{pa}_\mathcal{D}(\tau)$ and τ. Furthermore, let

$$b_\tau = (\beta_{ij} \mid i \in \tau, j \in \mathrm{pa}_\mathcal{D}(\tau), j \to i) \in \mathbb{R}^{d_\tau} \tag{3.16}$$

be the column vector comprising all non-vanishing components of β_τ. In order to compute the derivative of $\ell(\beta_\tau, \Lambda_\tau)$ with respect to b_τ, we use the vec-operator of column-wise vectorization to rewrite

$$\mathrm{tr}(\Lambda_\tau^{-1}\beta_\tau S_{\mathrm{pa}_\mathcal{D}(\tau),\tau}) = \mathrm{vec}(\Lambda_\tau^{-1}S_{\tau,\mathrm{pa}_\mathcal{D}(\tau)})'\mathrm{vec}(\beta_\tau), \tag{3.17}$$

compare Harville (1997, eqn. (16.2.14)). Furthermore, using Harville (1997, eqn. (16.2.15)),

$$\mathrm{tr}(\Lambda_\tau^{-1}\beta_\tau S_{\mathrm{pa}_\mathcal{D}(\tau),\mathrm{pa}_\mathcal{D}(\tau)}\beta_\tau') = \mathrm{vec}(\beta_\tau)'(S_{\mathrm{pa}_\mathcal{D}(\tau),\mathrm{pa}_\mathcal{D}(\tau)} \otimes \Lambda_\tau^{-1})\mathrm{vec}(\beta_\tau). \tag{3.18}$$

Now let $i \in \tau$, $j \in \mathrm{pa}_\mathcal{D}(\tau)$ be such that $j \to i$. Then by plugging the expressions from (3.17) and (3.18) into (3.12), we obtain that

$$\frac{\partial \ell(\beta_\tau, \Lambda_\tau)}{\partial \beta_{ij}} = -\frac{n}{2}\left[2 \times \frac{\partial}{\partial \beta_{ij}}\{\mathrm{vec}(\Lambda_\tau^{-1}S_{\tau,\mathrm{pa}_\mathcal{D}(\tau)})'\mathrm{vec}(\beta_\tau)\} - \frac{\partial}{\partial \beta_{ij}}\{\mathrm{vec}(\beta_\tau)'(S_{\mathrm{pa}_\mathcal{D}(\tau),\mathrm{pa}_\mathcal{D}(\tau)} \otimes \Lambda_\tau^{-1})\mathrm{vec}(\beta_\tau)\}\right]. \tag{3.19}$$

It follows that

$$\frac{\partial \ell(\beta_\tau, \Lambda_\tau)}{\partial b_\tau} = -n\left[\mathrm{vec}(\Lambda_\tau^{-1}S_{\tau,\mathrm{pa}_\mathcal{D}(\tau)})' - \mathrm{vec}(\beta_\tau)'(S_{\mathrm{pa}_\mathcal{D}(\tau),\mathrm{pa}_\mathcal{D}(\tau)} \otimes \Lambda_\tau^{-1})\right]\frac{\partial \mathrm{vec}(\beta_\tau)}{\partial b_\tau}, \tag{3.20}$$

which is a row vector with d_τ components. Let π_τ be the derivative matrix

$$\pi_\tau = \frac{\partial \mathrm{vec}(\beta_\tau)}{\partial b_\tau} \in \mathbb{R}^{[\tau \times \mathrm{pa}_{\mathcal{D}}(\tau)] \times d_\tau}, \tag{3.21}$$

which does not depend on b_τ and has entries in $\{0, 1\}$. This matrix clearly satisfies

$$\mathrm{vec}(\beta_\tau) = \pi_\tau b_\tau, \quad \forall\, b_\tau \in \mathbb{R}^{d_\tau}, \tag{3.22}$$

and in fact, each row of π_τ has precisely one entry equal to one and the remaining entries equal to zero. From (3.20), (3.21), and (3.22), we obtain that

$$\frac{\partial \ell(\beta_\tau, \Lambda_\tau)}{\partial b_\tau} = -n \left[\mathrm{vec}(\Lambda_\tau^{-1} S_{\tau, \mathrm{pa}_{\mathcal{D}}(\tau)})' - b_\tau' \pi_\tau' (S_{\mathrm{pa}_{\mathcal{D}}(\tau), \mathrm{pa}_{\mathcal{D}}(\tau)} \otimes \Lambda_\tau^{-1}) \right] \pi_\tau. \tag{3.23}$$

Setting the derivatives in (3.14) and (3.23) equal to zero, we obtain the likelihood equations

$$b_\tau = \left[\pi_\tau' (S_{\mathrm{pa}_{\mathcal{D}}(\tau), \mathrm{pa}_{\mathcal{D}}(\tau)} \otimes \Lambda_\tau^{-1}) \pi_\tau \right]^{-1} \pi_\tau' \mathrm{vec}(\Lambda_\tau^{-1} S_{\tau, \mathrm{pa}_{\mathcal{D}}(\tau)}), \tag{3.24}$$

and

$$(\Lambda_\tau)_{ij} = S(\beta_\tau)_{ij}, \quad \forall\, i, j \in \tau,\ i - j \text{ in } G. \tag{3.25}$$

3.5 *Maximum likelihood estimation by generalized least squares and iterative proportional fitting*

The condition $n \geq p$ implies the almost sure positive definiteness of the sample covariance matrix S from (3.3). A positive definite sample covariance matrix implies in turn the existence of the global maximum of $\ell(\Sigma)$ over $\mathbf{P}(G)$. This condition is not necessary in general but we are not aware of any results in the literature which provide a necessary and sufficient condition. In general, the likelihood function $\ell(\beta_\tau, \Lambda_\tau)$ may exhibit more than one local maximum, and the likelihood equations (3.24) and (3.25) may admit more than one solution; compare Chapter 2. Here, we propose a simple iterative algorithm, which generates a sequence of estimates of $(\beta_\tau, \Lambda_\tau)$, whose accumulation points solve the likelihood equations. This algorithm is immediately suggested by the form of the likelihood equations.

First, consider the problem of solving equation (3.25) while fixing the mean parameters $\beta_\tau \in \mathbf{B}_\tau(G)$. Clearly, (3.25) constitutes the likelihood equations for an undirected graph

model for G_τ with the synthetic sample covariance matrix $S(\beta_\tau)$, which is the sample covariance matrix of the residuals in the block-regression (3.7). It follows from Lauritzen (1996, Thm. 5.3) that for fixed β_τ there exists a unique solution $\hat{\Lambda}_\tau(\beta_\tau) \in \mathbf{P}(G_\tau)$ to equation (3.25), which can be computed by the iterative proportional fitting algorithm (Whittaker, 1990, pp.182–185). Moreover, $\hat{\Lambda}_\tau(\beta_\tau)$ is the unique maximizer in the maximization problem

$$\max\{\ell(\beta_\tau, \Lambda_\tau) \mid \Lambda_\tau \in \mathbf{P}(G_\tau)\}. \tag{3.26}$$

Second, it is obvious that equation (3.24), which is in a generalized least squares form, admits a unique solution for fixed $\Lambda_\tau \in \mathbf{P}(G_\tau)$. We denote this solution by $\hat{b}_\tau(\Lambda_\tau)$ and let

$$\hat{\beta}_\tau(\Lambda_\tau) = \pi_\tau \hat{b}_\tau(\Lambda_\tau). \tag{3.27}$$

Clearly, $\hat{\beta}_\tau(\Lambda_\tau)$ is the unique maximizer in the maximization problem

$$\max\{\ell(\beta_\tau, \Lambda_\tau) \mid \beta_\tau \in \mathbf{B}_\tau(G)\}. \tag{3.28}$$

These partial maximization properties suggest to solve the likelihood equations by choosing a starting value $\beta_\tau^{(0)} \in \mathbf{B}_\tau(G)$ or $\Lambda_\tau^{(0)} \in \mathbf{P}(G_\tau)$ and iteratively computing

$$\begin{aligned} \Lambda_\tau^{(i+1)} &= \hat{\Lambda}_\tau(\beta_\tau^{(i)}), \\ \beta_\tau^{(i+1)} &= \hat{\beta}_\tau(\Lambda_\tau^{(i)}). \end{aligned} \tag{3.29}$$

A simple starting value $\Lambda_\tau^{(0)} \in \mathbf{P}(G_\tau)$ is the identity matrix I_τ. The iterative algorithm in (3.29) is a partial maximization algorithm based on maximizations over sections in the parameter space that admit unique solutions. Thus, whenever the data are such that the sample covariance matrix is positive definite and that there are only finitely many solutions to the likelihood equations that fall in the same contour of the likelihood, then Corollary A.5 in Appendix A implies that, regardless of the choice of a feasible starting value, the algorithm in (3.29) generates a sequence of estimates $(\Lambda_\tau^{(i)}, \beta_\tau^{(i)})$ that converges to a local maximum or saddle point of the likelihood function.

3.6 Conclusion

We have described an algorithm for solving the likelihood equations in Gaussian AMP chain graph models. This algorithm is a combination of generalized least squares and iterative

proportional fitting. As an iterative partial maximization algorithm it is guaranteed to converge to a solution of the likelihood equations if the data are such that there exist only finitely many solutions to the likelihood equations at which the likelihood takes on the same value.

Chapter 4

ITERATIVE CONDITIONAL FITTING FOR COVARIANCE GRAPH MODELS

Graphical models with bidirected edges ($\leftrightarrow$) represent marginal independence: the absence of an edge between two vertices indicates that the corresponding variables are marginally independent. In this chapter, we consider maximum likelihood estimation in the case of continuous variables with a Gaussian joint distribution, sometimes termed a covariance graph model. We present a new fitting algorithm which exploits standard regression techniques and establish its convergence properties. Moreover, we contrast our procedure to existing estimation algorithms. We give the new algorithm the name Iterative Conditional Fitting since in each step of the procedure, a conditional distribution is estimated, subject to constraints, while a marginal distribution is held fixed. This approach is in duality to the well-known iterative proportional fitting algorithm, in which marginal distributions are fitted while conditional distributions are held fixed.

4.1 Introduction

Graphical models are commonly based on undirected graphs, DAGs, or chain graphs in which the absence of an edge between two vertices indicates some conditional independence between the associated variables. However, there has also been interest in graphical models in which the variables associated with two non-adjacent vertices are marginally independent. These models for marginal independence may naturally be represented with bidirected edges ($\leftrightarrow$) via a natural extension of d-separation. These models correspond to the special case of *ancestral graph models* (Richardson and Spirtes, 2002) where the ancestral graph has bidirected edges only. The case of jointly Gaussian variables has also been termed *covariance graph* models by Cox and Wermuth (1993, 1996) who use dashed lines rather than

bidirected edges. Besides being interesting models in their own right, graphical models for marginal independence are also interesting in the context of DAGs since certain DAGs with hidden variables induce marginal independences amongst the observed variables that can be represented by a bidirected graph (see §4.2).

For undirected graphs, DAGs and chain graphs, parameter learning procedures are well developed, see e.g. Lauritzen (1996) or Whittaker (1990), and many methods are implemented in the software package MIM (Edwards, 2000). This is not the case, however, for graphical models for marginal independence and covariance graph models in particular. For instance, MIM does not permit maximum likelihood (ML) estimation in covariance graph models but permits fitting only by a heuristic method due to Kauermann (1996), which is based on a "dual likelihood"; see also Edwards (2000, §7.4) and Banerjee and Richardson (2003).

This chapter presents a new iterative algorithm for ML estimation in covariance graph models. In §4.2 we describe and motivate graphical models for marginal independence in general, and in §4.3 we turn to the Gaussian case of covariance graph models. In particular, we review and critique Anderson's ML estimation algorithm. In §4.4 we present our new algorithm which converges to a solution of the likelihood equations for almost every value of the observations. In §4.5 we show the estimates for an example data set and in §4.6 we outline future extensions to our algorithm and comment on related literature.

4.2 Graphical models for marginal independence

Suppose that we observe the set of variables V in the random vector $X = (X_i \mid i \in V)$. Let $G = (V, E)$ be a graph with the variable set V as vertex set and the edge set $E \subsetneq V \times V$ consisting exclusively of bidirected edges (i, j), $i \neq j$, $i, j \in V$, denoted by $i \leftrightarrow j$.

4.2.1 Global Markov property for bidirected graphs

In the bidirected graph G, a path between vertices $i, j \in V$ is said to *m-connect* given $S \subseteq V$ if there is a path between i and j on which every non-endpoint vertex is in S. Disjoint non-empty sets of vertices A and B are *m-connected* given S in G if for some $i \in A$ and $j \in B$

there exists an m-connecting path between i and j given S in G. Otherwise, A and B are *m-separated* given S where S is allowed to be empty. The distribution of X satisfies the *global Markov property* for G if $X_A \perp\!\!\!\perp X_B \mid X_S$ holds whenever A is m-separated from B given S. Here, $X_A = (X_i \mid i \in A)$, etc. Note that the global Markov property implies the *pairwise Markov property* consisting of the marginal independences $X_i \perp\!\!\!\perp X_j$ for all $i \not\leftrightarrow j$. If X has a Gaussian distribution then the pairwise Markov property also implies the global Markov property but this does not need to hold for an arbitrary probability distribution.

In the graph shown in Figure 4.1(a), the path $1 \leftrightarrow 3 \leftrightarrow 4 \leftrightarrow 2$ m-connects 1 and 2 given $Z = \{3, 4\}$, but 1 and 2 are m-separated given any proper subset of $\{3, 4\}$. The global Markov property for this graph requires that $1 \perp\!\!\!\perp (2, 4)$ and $2 \perp\!\!\!\perp (1, 3)$. Note that these conditional independences are also stated by the AMP Markov property for the chain graph in Figure 1.1(c).

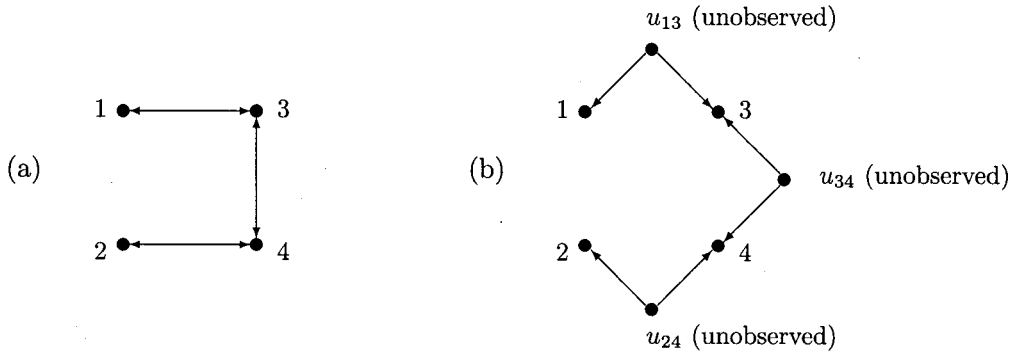


Figure 4.1: (a) Bidirected graph. (b) DAG with hidden variables u_{13}, u_{34}, u_{24} (§4.2.3).

In a bidirected graph, there are no directed paths, and every non-endpoint vertex on a path is a *collider*, i.e. two arrow-heads point to the vertex. Therefore, the definition of m-connection given above is the natural extension of Pearl's (1988) d-connection criterion to graphs containing bidirected edges. Richardson and Spirtes (2002) define m-connection for a larger class of graphs called ancestral graphs, which may contain directed, undirected and bidirected edges. The definition for ancestral graphs reduces to the definition given

here in the case where all edges are bidirected (compare Chapter 5).

4.2.2 Connected set Markov property

A set C in a bidirected graph G is said to be *connected* if for every pair of vertices $i, j \in C$ there is a path between i and j involving only vertices in C. Let the *spouses* of $i \in V$ be $\mathrm{sp}(i) = \{j \mid i \leftrightarrow j\}$ and the *non-spouses* $\mathrm{nsp}(i) = V \setminus (\mathrm{sp}(i) \cup \{i\})$. For a set $C \subseteq V$, we define $\mathrm{sp}(C) = \cup_{i \in C} \mathrm{sp}(i)$ and $\mathrm{nsp}(C) = \cap_{i \in C} \mathrm{nsp}(i)$. Then the distribution of X satisfies the *connected set Markov property* for G if $X_C \perp\!\!\!\perp X_{\mathrm{nsp}(C)}$ for every connected set C. In words, every connected set is marginally independent of all other variables other than those adjacent to some element in the set. Richardson (2003) proves that an arbitrary probability distribution obeys the connected set Markov property iff it obeys the global Markov property. The connected set Markov property for the graph in Figure 4.1(a) requires: $1 \perp\!\!\!\perp (2,4)$, $2 \perp\!\!\!\perp (1,3)$.

4.2.3 Relation to DAG models with hidden variables

Graphical models for marginal independence can be motivated by the following consideration (see also Pearl and Wermuth, 1994). Suppose that there is DAG D with vertex set $V \cup U$, where the variables in V are observed, and those in U are unobserved. Suppose further that observed variables $i \in V$ have no children in the graph, i.e. $\mathrm{ch}_D(i) = \{j \in V \cup U \mid i \to j\} = \emptyset$. Models of this kind are used in psychology and the social sciences (see e.g. Bollen, 1989, §6).

From the DAG D, construct the induced bidirected graph $G(D)$ over V by including the bidirected edge $i \leftrightarrow j$, $i, j \in V$, if $\mathrm{an}_D(i) \cap \mathrm{an}_D(j) \neq \emptyset$, where $\mathrm{an}_D(i) = \{j \in V \mid j \to \ldots \to i \text{ or } j = i\}$ are the ancestors of i in D. Note that $\mathrm{an}_D(i) \cap \mathrm{an}_D(j) \subseteq U$. It then follows as a special case of Theorem 4.18 in Richardson and Spirtes (2002) that $G(D)$ represents the Markov structure induced by D on the observed variables.

Proposition 4.1. *Let D be a DAG with vertex set $V \cup U$ such that $\mathrm{ch}_D(i) = \emptyset$ for all $i \in V$, and let $G(D)$ be the bidirected graph with vertex set V defined above. Then for any*

disjoint sets $A, B, S \subseteq V$, *with* $A, B \neq \emptyset$,

$$A \text{ and } B \text{ are d-separated given } S \text{ in } D$$
$$\Leftrightarrow A \text{ and } B \text{ are m-separated given } S \text{ in } G(D).$$

Figure 4.1(b) shows a DAG D with $G(D)$ equal to the graph shown in Figure 4.1(a).

DAGs with hidden variables that satisfy the conditions of Proposition 4.1 induce an independence structure over the observed variables that can be represented by a bidirected graph. However, further (non-independence) restrictions can hold in the hidden variable model such that it is only a submodel of the bidirected graph model (Richardson and Spirtes, 2002, §7.3.1, §8.6).

4.3 Covariance graph models

Suppose now that the variables V are continuous with a centered Gaussian $\equiv$ normal joint distribution, i.e. $X \sim \mathcal{N}_V(0, \Sigma) \in \mathbb{R}^V$ where $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{V \times V}$ is the unknown positive definite covariance matrix. The normal distribution of X is pairwise and globally Markov for the bidirected graph $G = (V, E)$ iff

$$X_i \perp\!\!\!\perp X_j \quad \Longleftrightarrow \quad \sigma_{ij} = 0 \quad \Longleftrightarrow \quad i \not\leftrightarrow j. \tag{4.1}$$

Let $\mathbf{P}(V)$ be the cone of all positive definite $V \times V$ matrices and let $\mathbf{P}(G)$ be the cone of all matrices $\Sigma \in \mathbf{P}(V)$ which fulfill the linear restrictions in (4.1). Then the *covariance graph model* based on G is the family of all normal distributions

$$\mathbf{N}(G) = \big(\mathcal{N}_V(0, \Sigma) \mid \Sigma \in \mathbf{P}(G)\big). \tag{4.2}$$

This chapter considers the estimation of the unknown parameter Σ based on a sample of i.i.d. observations $X^{(k)}$, $k \in N = \{1, \ldots, n\}$, from the covariance graph model (4.2). The set N can be interpreted as indexing the subjects on which we observe the variables in V. We group the vectors in the sample as columns in the $V \times N$ random matrix Y which is distributed as

$$Y \in \mathbb{R}^{V \times N} \sim \mathcal{N}_{V \times N}(0, \Sigma \otimes I_N). \tag{4.3}$$

Here, I_N is the $N \times N$ identity matrix, $\Sigma \in \mathbf{P}(G)$ is the unknown positive definite covariance matrix, and $\otimes$ is the Kronecker product. Thus the i-th row $Y_i = Y_{i\cdot} \in \mathbb{R}^N$ of the matrix Y contains the i.i.d. observations for variable $i \in V$ on all the subjects in N and the k-th column $Y_{\cdot k} = X^{(k)}$ holds all the observations made on the subject $k \in N$. Finally, the sample size is $n = |N|$ and the number of variables is $p = |V|$.

Since our model assumes a zero mean, the empirical covariance matrix is defined to be

$$S = \frac{1}{n} YY' \in \mathbb{R}^{V \times V}. \tag{4.4}$$

We shall assume that $n \geq p$ such that S is positive definite with probability one.

Note that the case where the model also includes an unknown mean vector μ can be treated by estimating μ by the empirical mean vector $\bar{Y} \in \mathbb{R}^V$, i.e., the vector of the row means of Y. The empirical covariance matrix would then be the matrix

$$\tilde{S} = \frac{1}{n}(Y - \bar{Y} \otimes 1_N)(Y - \bar{Y} \otimes 1_N)' \in \mathbb{R}^{V \times V}, \tag{4.5}$$

where $1_N = (1, \ldots, 1) \in \mathbb{R}^N$. However, we would have to assume that $n \geq p + 1$ to ensure that $\tilde{S}$ is positive definite with probability one.

4.3.1 Maximum likelihood estimation

The density function of Y with respect to the Lebesgue measure is the function $f_\Sigma : \mathbb{R}^{V \times N} \to \mathbb{R}$ given by

$$f_\Sigma(y) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\Big\{ -\frac{1}{2} \sum_{k \in N} y'_{\cdot k} \Sigma^{-1} y_{\cdot k} \Big\}. \tag{4.6}$$

Considered as a function of the unknown parameters for fixed data y this gives the likelihood function of the covariance graph model $\mathbf{N}(G)$ as the mapping $L : \mathbf{P}(G) \to \mathbb{R}$ where $L(\Sigma) = f_\Sigma(y)$. In ML estimation, the parameter Σ is estimated by the element of $\mathbf{P}(G)$ which maximizes the likelihood L or more conveniently the log-likelihood $\ell(\Sigma) = \log L(\Sigma)$.

The log-likelihood $\ell(\Sigma)$ can be expressed as

$$\ell(\Sigma) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{n}{2} \mathrm{tr}\{\Sigma^{-1} S\}, \tag{4.7}$$

see e.g. Edwards (2000, §3.1). The condition $n \geq p$ implies the almost sure existence of the global maximum of $\ell(\Sigma)$ over $\mathbf{P}(G)$. This condition is not necessary in general but we are not aware of any results in the literature which provide a necessary and sufficient condition.

Besides existence, there is also the issue of uniqueness of the ML estimates, i.e. whether the likelihood has a unique local maximum which is then global. The model $\mathbf{N}(G)$ is a curved but not regular exponential family, thus, the log-likelihood need not be concave. In fact, the log-likelihood can have multiple local maxima (compare Chapter 2). For a large enough sample size, a multimodal likelihood seems not to arise often in practice assuming the model assumptions hold but might still arise if the model assumptions do not hold (see also Cox and Wermuth, 1996, p. 102f).

4.3.2 The likelihood equations

The likelihood equations are the estimating equations obtained by setting the derivatives of the log-likelihood $\ell(\Sigma)$ with respect to σ_{ij}, $i = j$ or $i \leftrightarrow j$, to zero. From Anderson and Olkin (1985, §2.1.1) it follows that the likelihood equations are

$$(\Sigma^{-1})_{ij} = (\Sigma^{-1} S \Sigma^{-1})_{ij} \tag{4.8}$$

for $i = j$ and $i \leftrightarrow j$. The full matrix Σ is then determined by $\sigma_{ij} = 0$ for $i \not\leftrightarrow j$.

4.3.3 Anderson's algorithm

Anderson (1969, 1970) studied general linear hypotheses on the covariance matrix Σ which contain covariance graph models as a special case. In Anderson (1973), he developed an iterative algorithm to solve specifically the likelihood equations under linear hypotheses on Σ. We refer to this estimation procedure as *Anderson's algorithm.*

The iterations in Anderson's algorithm consist of solving a system of linear equations built from the current estimate of Σ. In the case of a covariance graph model, the linear equations are solved for the vector of unrestricted elements in Σ, i.e. for $\sigma = (\sigma_{ij} \mid (i,j) \in F)$ where $F = \{ij \equiv (i,j) \mid i \leftrightarrow j \vee i = j\}$, and the algorithm can be specified as follows. We

denote $\sigma^{ij} = (\Sigma^{-1})_{ij}$ and define $A = A_\Sigma$ to be the $F \times F$ matrix with

$$A_{(ij,kk)} = \sigma^{ik}\sigma^{jk} \qquad \forall ij \in F,\ k \in V, \tag{4.9}$$

$$A_{(ij,k\ell)} = \sigma^{ik}\sigma^{j\ell} + \sigma^{jk}\sigma^{i\ell} \qquad \forall ij \in F,\ k \leftrightarrow \ell. \tag{4.10}$$

Furthermore, we set the $F \times 1$ vector $b = b_\Sigma$ to

$$b_{ij} = (\Sigma^{-1} S \Sigma^{-1})_{ij} \quad \forall ij \in F. \tag{4.11}$$

It can be shown that $\Sigma \in \mathbf{P}(G)$ solves $A_\Sigma \sigma = b_\Sigma$ iff Σ solves the likelihood equations (4.8).

Anderson's algorithm updates the current estimate $\Sigma^{(r)}$ to $\Sigma^{(r+1)}$ determined by the linear equations

$$A_{\Sigma^{(r)}}\ \sigma^{(r+1)} = b_{\Sigma^{(r)}}. \tag{4.12}$$

Thus, a fixed point of Anderson's algorithm is a solution to the likelihood equations (4.8). As starting value, Anderson suggests the identity matrix, i.e. $\Sigma^{(0)} = I_V$. In the first step, his algorithm constructs the empirical estimate $\Sigma^{(1)}$ with $\sigma_{ij}^{(1)} = S_{ij}$, $ij \in F$. However, neither $\Sigma^{(1)}$ nor any subsequent estimate of Σ has to be positive (semi-) definite and thus may not be a valid covariance matrix. Moreover, at any given stage, the likelihood may decrease, and convergence of Anderson's algorithm cannot be guaranteed.

Therefore, we propose an alternative algorithm for ML estimation which constructs a sequence of estimates in $\mathbf{P}(G)$ with never decreasing likelihood which converges to a solution of the likelihood equations for almost every data matrix y. Note that our new algorithm only fits covariance graph models and cannot treat the wide range of models covered by Anderson's algorithm.

4.4 *Iterative conditional fitting*

4.4.1 The idea

Let i be a variable index in V and set $-i = V \setminus \{i\}$. For $A, B \subseteq V$, $\Sigma_{A,B}$ denotes the $A \times B$ submatrix of Σ and Y_A denotes the $A \times N$ submatrix of Y. The conditional distribution of Y_i given Y_{-i} is the normal distribution

$$(Y_i \mid Y_{-i}) \sim \mathcal{N}_{\{i\}\times N}(B_i Y_{-i}, \lambda_i I_N) \in \mathbb{R}^{\{i\}\times N}, \tag{4.13}$$

where

$$B_i = \Sigma_{i,-i}\Sigma_{-i,-i}^{-1} \in \mathbb{R}^{\{i\}\times -i} \tag{4.14}$$

is the $\{i\} \times -i$ matrix of regression coefficients and

$$\lambda_i = \sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i} \in \mathbb{R} \tag{4.15}$$

is the conditional variance. The joint density function for Y can be factored into the product of the marginal density function for Y_{-i} and the conditional density function of Y_i given Y_{-i}:

$$f_\Sigma(y) = f_{(B_i,\lambda_i)}(y_i \mid y_{-i}) f_{\Sigma_{-i,-i}}(y_{-i}). \tag{4.16}$$

Our idea for an iterative ML estimation algorithm is then to treat $\Sigma_{-i,-i}$ as if it were known from a current feasible estimate of Σ, to estimate B_i and λ_i from the regression (4.13), and to then update Σ according to (4.14) and (4.15) using the three pieces $\Sigma_{-i,-i}$, B_i, and λ_i. Of course, this step needs to be carried out in turn for all $i \in V$.

The subtlety with this idea is that we need to respect the restriction $\Sigma \in \mathbf{P}(G)$ when estimating B_i and λ_i. If the graph G was the complete graph $\bar{G}$ in which an edge joins any pair of vertices then the mapping

$$\begin{aligned} \mathbf{P}(\bar{G}) = \mathbf{P}(V) &\to (0,\infty) \times \mathbb{R}^{\{i\}\times -i} \times \mathbf{P}(-i), \\ \Sigma &\mapsto (\lambda_i, B_i, \Sigma_{-i,-i}) \end{aligned} \tag{4.17}$$

would be bijective and the regression in (4.13) a standard normal regression. For a general graph G, this will not be the case. Fortunately, our covariance restrictions are linear and so simple that we will be able to find equivalent restrictions on the regression coefficients B_i which will lead to an equivalent standard normal regression based, however, on altered variables which we name *pseudo-variables.*

4.4.2 An example

Before we develop our idea to a generally applicable algorithm, we illustrate it by the example of the graph shown in Figure 4.1(a). This graph imposes the marginal independences

$1 \perp\!\!\!\perp (2,4)$ and $2 \perp\!\!\!\perp (1,3)$ which imply the conditional independences $3 \perp\!\!\!\perp 2 \mid 1$ and $4 \perp\!\!\!\perp 1 \mid 2$. Thus the joint density can be decomposed as

$$f(y) = f(y_3, y_4 \mid y_1, y_2) f(y_2) f(y_1). \tag{4.18}$$

In this factorization the term $f(y_3, y_4 \mid y_1, y_2)$ corresponds to the bivariate seemingly unrelated regression model considered in Chapter 2.

Here, however, we do not want to make use of the factorization (4.18) which holds in this special example but we wish to demonstrate how our algorithm proceeds in the general setting. Because of the symmetry of the graph we only describe the regressions $(Y_1 \mid Y_2, Y_3, Y_4)$ and $(Y_3 \mid Y_1, Y_2, Y_4)$. The remaining two regressions for Y_2 and Y_4 are the obvious analogs obtained by exchanging the indices 1 and 2, as well as 3 and 4.

Update Step For Variable 1

Let the *spouses* of $i \in V$ be $\mathrm{sp}(i) = \{j \mid i \leftrightarrow j\}$ and the *non-spouses* $\mathrm{nsp}(i) = V \setminus (\mathrm{sp}(i) \cup \{i\})$. For $i = 1$, we find $\mathrm{sp}(1) = 3$ and $\mathrm{nsp}(1) = 24$ where the shorthand ij denotes the set $\{i, j\}$. The bijection (4.17) suggests that we can improve our current estimate of Σ by holding the block $\hat{\Sigma}_{234,234}$ fixed and using the regression (4.13) to find improved estimates of B_i, λ_i and hence of $\Sigma_{1,1234} = \Sigma'_{1234,1}$. However, we need to respect the two restrictions $\sigma_{12} = \sigma_{14} = 0$. Since these are restrictions of marginal independence they do not translate into restricting some of the regression coefficients in B_1 to zero (as would be the case with an undirected graphical model). Instead some coefficients in B_1 will be a linear combination of the remaining entries in B_1.

More specifically, let β_{ij} and $\beta_{ij.K}$ denote the regression coefficient for Y_j in the regressions $(Y_i \mid Y_j)$ and $(Y_i \mid Y_{\{j\} \cup K})$, respectively. It follows from (4.14) that

$$B_1 \Sigma_{234,234} = \Sigma_{1,234} = (0, \sigma_{13}, 0). \tag{4.19}$$

Since $B_1 = (\beta_{12.34}, \beta_{13.24}, \beta_{14.23})$ this implies

$$\beta_{13.24} \Sigma_{3,24} + (\beta_{12.34}, \beta_{14.23}) \Sigma_{24,24} = (0, 0). \tag{4.20}$$

Thus,

$$\begin{aligned}(\beta_{12.34}, \beta_{14.23}) &= -\beta_{13.24}\Sigma_{3,24}\Sigma_{24,24}^{-1} \\ &= -\beta_{13.24}(\beta_{32.4}, \beta_{34.2}).\end{aligned} \tag{4.21}$$

From (4.21), it follows that $B_1 Y_{234} = \beta_{13.24} Z_1$ where the pseudo-variable Z_1 equals

$$Z_1 = Y_3 - \beta_{32.4} Y_2 - \beta_{34.2} Y_4 \in \mathbb{R}^{\{1\}\times N}. \tag{4.22}$$

The regression (4.13) then becomes

$$(Y_1 \mid Y_2, Y_3, Y_4) \sim \mathcal{N}(\beta_{13.24} Z_1, \lambda_1 I_N). \tag{4.23}$$

Since we hold $\hat{\Sigma}_{234,234}$ fixed it can be used to compute current estimates of the regression coefficients $\hat{\beta}_{32.4}$ and $\hat{\beta}_{34.2}$ which are plugged into (4.22) to yield the estimate $\hat{Z}_1$ of Z_1. We use $\hat{Z}_1$ in the regression (4.23) and find from the usual least squares formulas:

$$\begin{aligned}\hat{\beta}_{13.24} &= Y_1 \hat{Z}_1' (\hat{Z}_1 \hat{Z}_1')^{-1}, \\ \hat{\lambda}_1 &= \frac{1}{n} Y_1 \big(I_N - \hat{Z}_1' (\hat{Z}_1 \hat{Z}_1')^{-1} \hat{Z}_1\big) Y_1'.\end{aligned} \tag{4.24}$$

Using (4.21) we can compute $\hat{\beta}_{12.34}$ and $\hat{\beta}_{14.23}$ to complete the estimate $\hat{B}_1$.

With the new estimates, we update $\hat{\Sigma}$ as follows. The block $\hat{\Sigma}_{234,234}$ remains unchanged. From (4.14) and (4.15), we obtain $\hat{\Sigma}_{1,234} = \hat{B}_1 \hat{\Sigma}_{234,234}$ and $\hat{\Sigma}_{234,1} = \hat{\Sigma}_{1,234}'$. Finally, we update $\hat{\sigma}_{11}$ to $\hat{\sigma}_{11} = \hat{\lambda}_1 + \hat{B}_1 \hat{\Sigma}_{234,1}$.

Update Step For Variable 3

For $i = 3$, $\mathrm{sp}(3) = 14$ and $\mathrm{nsp}(3) = 2$. In regression (4.13), we must now respect $\sigma_{32} = 0$ which, by a similar calculation as in §4.4.2, translates into

$$\beta_{32.14} = -\beta_{31.24}\beta_{12} - \beta_{34.12}\beta_{42}. \tag{4.25}$$

Therefore the regression (4.13) is now

$$(Y_3 \mid Y_1, Y_2, Y_4) \sim \mathcal{N}\big((\beta_{31.24}, \beta_{34.12}) Z_3, \lambda_3 I_N\big) \tag{4.26}$$

with the pseudo-variables

$$Z_3 = \begin{pmatrix} Y_1 - \beta_{12} Y_2 \\ Y_4 - \beta_{42} Y_2 \end{pmatrix} \in \mathbb{R}^{\{1,4\} \times N}. \tag{4.27}$$

We fix $\hat{\Sigma}_{124,124}$ and compute from it the regression coefficients $\hat{\beta}_{12}$ and $\hat{\beta}_{42}$ yielding $\hat{Z}_3$ by (4.27). Using $\hat{Z}_3$ in the regression (4.26), we obtain the least squares estimates

$$\begin{aligned} \left(\hat{\beta}_{31.24}, \hat{\beta}_{34.12}\right) &= Y_3 \hat{Z}_3' (\hat{Z}_3 \hat{Z}_3')^{-1}, \\ \hat{\lambda}_3 &= \frac{1}{n} Y_3 \big(I_N - \hat{Z}_3' (\hat{Z}_3 \hat{Z}_3')^{-1} \hat{Z}_3\big) Y_3'. \end{aligned} \tag{4.28}$$

Using (4.25) we can compute $\hat{\beta}_{32.14}$ to complete the estimate $\hat{B}_3$. In the resulting update of $\hat{\Sigma}$, the submatrix $\hat{\Sigma}_{124,124}$ remains unchanged but we set $\hat{\Sigma}_{3,124} = \hat{B}_3 \hat{\Sigma}_{124,124}$ and $\hat{\Sigma}_{124,3} = \hat{\Sigma}_{3,124}'$. The remaining variance $\hat{\sigma}_{33}$ is updated to $\hat{\sigma}_{33} = \hat{\lambda}_3 + \hat{B}_3 \hat{\Sigma}_{124,3}$.

The Iteration

Figure 4.2 illustrates one full iteration in our algorithm in this example.

The algorithm cycles in arbitrary order through the four regressions $(Y_i \mid Y_{-i})$, $i = 1, 2, 3, 4$. In Figure 4.2, a filled circle represents variables in the conditioning set $-i$, and an unfilled circle stands for the variable i forming the response variable in the considered regression. The thick directed edges coincide with bidirected edges in the original graph shown in Figure 4.1. Thin edges do not have a corresponding bidirected edge in the original graph. Regression coefficients label the edges. It can be seen that the regression coefficients at thin edges are linear combinations of the regression coefficients at thick edges where the weights in the linear combinations are decorated with "hats" as $\hat{\beta}$ to remind the reader that they are computed from $\hat{\Sigma}_{-i,-i}$, the block remaining unchanged in the i-step of the iteration. Regression coefficients without "hat" are estimated by regression on appropriate pseudo-variables as illustrated in Figure 4.3.

A Criticism

Our algorithm does not make use of any available likelihood factorization. For example, (4.18) implies that the components $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$ of a solution to the likelihood equations

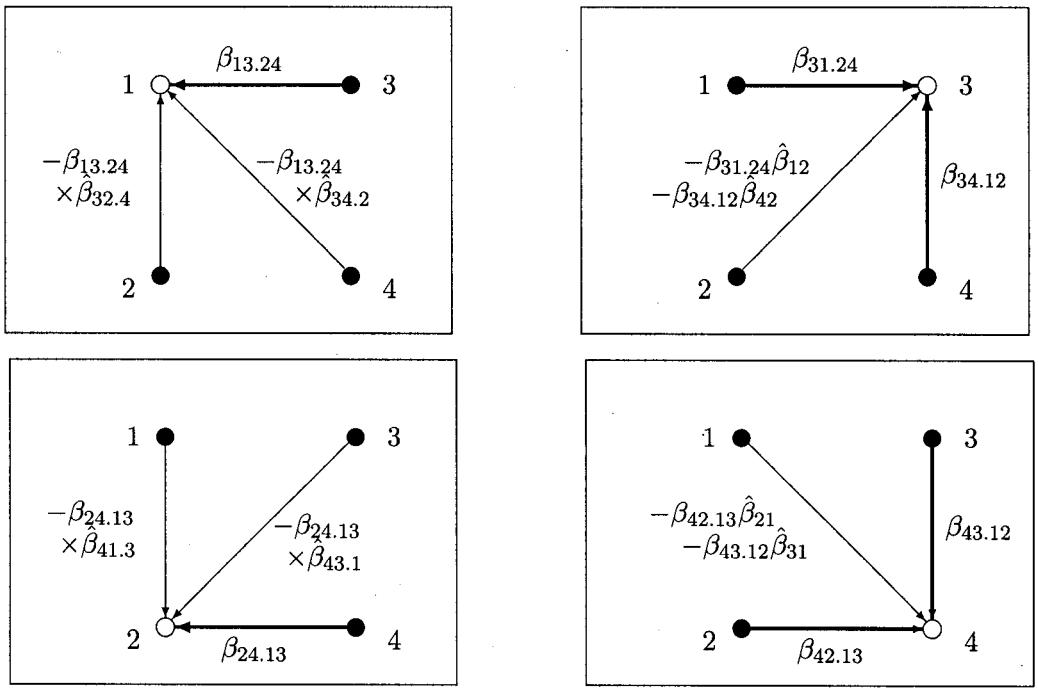


Figure 4.2: Illustration of new algorithm.

must coincide with the corresponding empirical quantities S_{11} and S_{22}, respectively. Thus the pseudo-variable regressions $(Y_1 \mid Y_2, Y_3, Y_4)$ and $(Y_2 \mid Y_1, Y_2, Y_4)$ need not be carried out, and the algorithm's convergence is sped up considerably. Hence, the algorithm may be improved by systematically employing information on which submatrices of a solution to the likelihood equations must coincide with their empirical counterparts, but this requires further work.

4.4.3 Pseudo-variable regression

We now describe the general algorithm. Let $\hat{\Sigma}^* \in \mathbf{P}(G)$ be a feasible estimate of Σ. Suppose we wish to update $\hat{\Sigma}^*$ to a new estimate $\hat{\Sigma} \in \mathbf{P}(G)$ by setting $\hat{\Sigma}_{-i,-i} = \hat{\Sigma}^*_{-i,-i}$ and using the regression $(Y_i \mid Y_{-i})$ to obtain $\hat{\Sigma}_{i,V}$. For $A \subseteq V$, define $\mathbf{P}_A(G)$ to be the set of all $A \times A$ submatrices of matrices in $\mathbf{P}(G)$. Then a matrix $\Sigma \in \mathbf{P}(G)$ iff Σ fulfills the two conditions

(1) $\Sigma_{-i,-i} \in \mathbf{P}_{-i}(G)$,

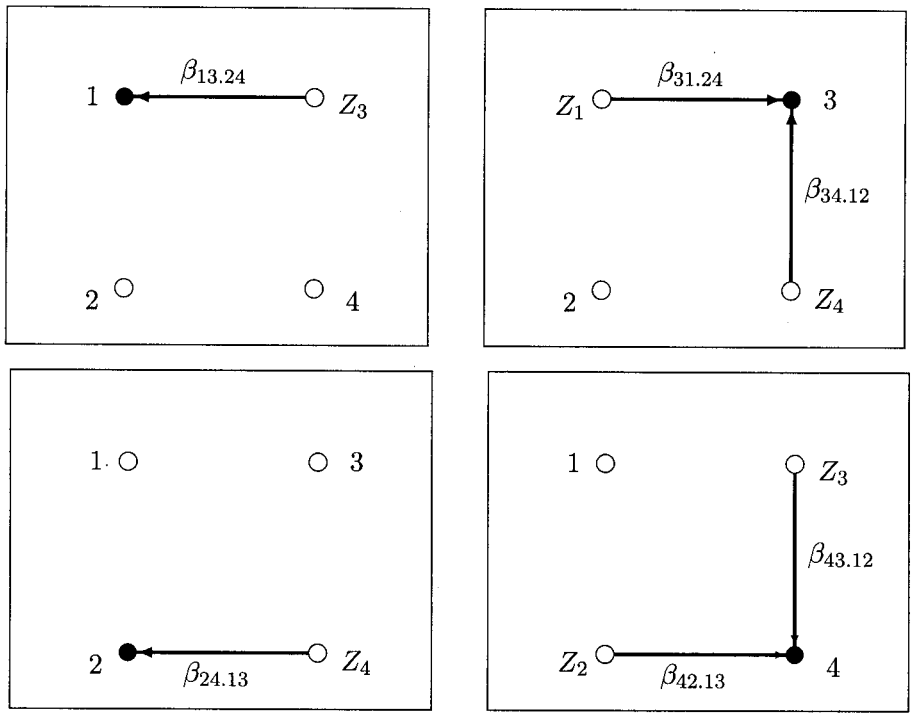


Figure 4.3: Illustration of the pseudo-variable regressions.

(2) $(B_i\Sigma_{-i,-i})_{ij} = \sigma_{ij} = 0 \quad \forall i \not\leftrightarrow j$; compare (4.14).

Since $\hat{\Sigma}_{-i,-i} = \hat{\Sigma}^*_{-i,-i}$ condition (1) is fulfilled by $\hat{\Sigma}$ because $\hat{\Sigma}^* \in \mathbf{P}(G)$ by assumption. The remaining condition (2) can be rewritten as $B_i\Sigma_{-i,\mathrm{nsp}(i)} = 0$ and further as

$$B_{i,\mathrm{sp}(i)}\Sigma_{\mathrm{sp}(i),\mathrm{nsp}(i)} + B_{i,\mathrm{nsp}(i)}\Sigma_{\mathrm{nsp}(i),\mathrm{nsp}(i)} = 0. \tag{4.29}$$

Hence, the regression coefficients for the non-spouses of i are linear combinations of the regression coefficients for the spouses:

$$\begin{aligned} B_{i,\mathrm{nsp}(i)} &= -B_{i,\mathrm{sp}(i)}\Sigma_{\mathrm{sp}(i),\mathrm{nsp}(i)}\Sigma^{-1}_{\mathrm{nsp}(i),\mathrm{nsp}(i)} \\ &= -B_{i,\mathrm{sp}(i)}B_{\mathrm{sp}(i).\mathrm{nsp}(i)}, \end{aligned} \tag{4.30}$$

where

$$B_{\mathrm{sp}(i).\mathrm{nsp}(i)} = \Sigma_{\mathrm{sp}(i),\mathrm{nsp}(i)}\Sigma^{-1}_{\mathrm{nsp}(i),\mathrm{nsp}(i)} \tag{4.31}$$

are the regression coefficients in $(Y_{\mathrm{sp}(i)} \mid Y_{\mathrm{nsp}(i)})$. From (4.30), we obtain that the mapping

$$\begin{aligned} \mathbf{P}(G) &\to (0,\infty) \times \mathbb{R}^{\{i\}\times \mathrm{sp}(i)} \times \mathbf{P}_{-i}(G), \\ \Sigma &\mapsto (\lambda_i, B_{i,\mathrm{sp}(i)}, \Sigma_{-i,-i}) \end{aligned} \tag{4.32}$$

is bijective. Hence, for restricted $\Sigma \in \mathbf{P}(G)$ the submatrix $\Sigma_{-i,-i}$ does not restrict the range of variation of λ_i and $B_{i,\mathrm{sp}(i)}$ in the maximization of the likelihood factor $f_{(B_i,\lambda_i)}(y_i \mid y_{-i})$ in (4.16). Moreover, for $\Sigma \in \mathbf{P}(G)$, the regression (4.13) can be rewritten as

$$(Y_i \mid Y_{-i}) \sim \mathcal{N}_{\{i\}\times N}(B_{i,\mathrm{sp}(i)} Z_i, \lambda_i I_N) \in \mathbb{R}^{\{i\}\times N}, \tag{4.33}$$

where the *pseudo-variables* Z_i are the residuals in the regression $(Y_{\mathrm{sp}(i)} \mid Y_{\mathrm{nsp}(i)})$, i.e.

$$Z_i = Y_{\mathrm{sp}(i)} - B_{\mathrm{sp}(i).\mathrm{nsp}(i)} Y_{\mathrm{nsp}(i)} \in \mathbb{R}^{\mathrm{sp}(i)\times N}. \tag{4.34}$$

Since $\mathrm{sp}(i) \subseteq -i$ and $\mathrm{nsp}(i) \subseteq -i$ the regression coefficients $\hat{B}_{\mathrm{sp}(i).\mathrm{nsp}(i)}$ can be calculated from $\hat{\Sigma}^*_{-i,-i}$. These can be plugged into (4.34) to find estimates $\hat{Z}_i$ of the pseudo-variables. The pseudo-variable regression (4.33) yields new estimates $\hat{B}_{i,\mathrm{sp}(i)}$ and $\hat{\lambda}_i$ from which we can calculate $\hat{B}_{i,\mathrm{nsp}(i)}$ using (4.30). Thus, we can form the estimate $\hat{B}_i$ from which we can reconstruct $\hat{\Sigma}_{i,-i}$ using (4.14) and $\hat{\sigma}_{ii}$ using (4.15).

4.4.4 The algorithm

These considerations lead to the following algorithm where $\hat{\Sigma}^{(r)}$ denotes the estimated value of Σ after the r-th iteration, and $\hat{\Sigma}^{(r,i)}$ is the estimated value of Σ after the i-th step of the r-th iteration, i.e. after regressing Y_i on Y_{-i}.

(1) Set the iteration counter $r = 0$, and choose a starting value $\hat{\Sigma}^{(0)} \in \mathbf{P}(G)$, e.g. the identity matrix $\hat{\Sigma}^{(0)} = I_V$.

(2) Order the variables in V as $V = \{1, \ldots, p\}$, set $\hat{\Sigma}^{(r,0)} = \hat{\Sigma}^{(r)}$, and repeat the following steps for all $i = 1, \ldots, p$:

   (a) Let $\hat{\Sigma}^{(r,i)}_{-i,-i} = \hat{\Sigma}^{(r,i-1)}_{-i,-i}$ and calculate from this submatrix the regression coefficients $\hat{B}_{\mathrm{sp}(i).\mathrm{nsp}(i)}$ according to (4.31). Construct the pseudo-variables $\hat{Z}_i$ by plugging $\hat{B}_{\mathrm{sp}(i).\mathrm{nsp}(i)}$ into (4.34).

(b) Compute the MLE of $B_{i,\mathrm{sp}(i)}$ and λ_i in the linear regression (4.33):

$$\begin{aligned} \hat{B}_{i,\mathrm{sp}(i)} &= Y_i \hat{Z}_i' (\hat{Z}_i \hat{Z}_i')^{-1}, \\ \hat{\lambda}_i &= \frac{1}{n} Y_i \big(I_N - \hat{Z}_i' (\hat{Z}_i \hat{Z}_i')^{-1} \hat{Z}_i\big) Y_i'. \end{aligned} \tag{4.35}$$

(c) Use (4.30) to compute $\hat{B}_{i,\mathrm{nsp}(i)}$ which completes $\hat{B}_i$. Inverting (4.14) and (4.15), reconstruct $\hat{\Sigma}^{(r,i)}_{i,-i} = \hat{B}_i \hat{\Sigma}^{(r,i)}_{-i,-i}$, set $\hat{\Sigma}^{(r,i)}_{-i,i}$ equal to the transpose of $\hat{\Sigma}^{(r,i)}_{i,-i}$, and complete $\hat{\Sigma}^{(r,i)}$ by setting $\hat{\sigma}^{(r,i)}_{ii} = \hat{\lambda}_i + \hat{B}_i \hat{\Sigma}^{(r,i)}_{-i,i}$.

(3) Set $\hat{\Sigma}^{(r+1)} = \hat{\Sigma}^{(r,p)}$. Increment the counter r to $r+1$. Go to (2).

The iterations can be stopped according to a criterion such as "the estimate of Σ is not changed" (in some pre-determined accuracy).

4.4.5 Convergence

The key to prove convergence properties of the algorithm in §4.4.4 is to recognize that the algorithm consists of iterated partial maximizations over sections of the parameter space $\mathbf{P}(G)$; compare Lauritzen (1996, Appendix A.4) and Appendix A. More accurately, we will consider the parameter space

$$\Theta = \{\Sigma \in \mathbf{P}(G) \mid \ell(\Sigma) \geq \ell(\hat{\Sigma}^{(0)})\} \tag{4.36}$$

which contains the global maximizer of $\ell(\Sigma)$. The set Θ is closed, and under the condition $n \geq p$ almost surely bounded. Thus, ignoring a null set of observations, Θ is compact.

The section $\Theta_i(\tilde{\Sigma}) \subsetneq \Theta$ is defined by

$$\Theta_i = \big\{\Sigma \in \Theta \mid \Sigma_{-i,-i} = \tilde{\Sigma}_{-i,-i}\big\}. \tag{4.37}$$

The bijection (4.32) implies that the algorithm steps (2a)-(2c) maximize the log-likelihood partially over the section $\Theta_i(\hat{\Sigma}^{(r,i-1)})$, i.e.

$$\hat{\Sigma}^{(r,i)} = \arg\max \big\{\ell(\Sigma) \mid \Sigma \in \Theta_i(\hat{\Sigma}^{(r,i-1)})\big\}. \tag{4.38}$$

Hence, the sequence $\ell(\hat{\Sigma}^{(r)})$ is non-decreasing and bounded and thus converges, i.e.

$$\lim_{r\to\infty} \ell(\hat{\Sigma}^{(r)}) = \ell^{(\infty)}. \tag{4.39}$$

Since Θ is compact the sequence $\hat{\Sigma}^{(r)}$ must have a convergent subsequence $\hat{\Sigma}^{(r_\iota)}$ with limit $\hat{\Sigma}^{(\infty)}$. By (4.38), $\hat{\Sigma}^{(\infty)}$ maximizes the log-likelihood in particular over every section of Θ defined by fixing all but one single entry σ_{ij} of Σ. This implies that $\hat{\Sigma}^{(\infty)}$ solves the likelihood equations. Moreover, (4.38) shows that $\hat{\Sigma}^{(\infty)}$ is either a saddle point or a local maximum of the log-likelihood. Finally, (4.39) implies that $\ell(\hat{\Sigma}^{(\infty)}) = \ell^{(\infty)}$.

The results in Appendix A yield that if the likelihood equations have only finitely many solutions then our algorithm converges to a local maximum or saddle point of the likelihood. The following theorem summarizes our results.

Theorem 4.2. *Suppose the sequence $(\hat{\Sigma}^{(r)})$ is constructed by the algorithm from §4.4.4. Then all accumulation points of $(\hat{\Sigma}^{(r)})$ are saddle points or local maxima of the log-likelihood. Moreover, all accumulation points have the same likelihood value. In particular, if the likelihood equations have only finitely many solutions, then $(\hat{\Sigma}^{(r)})$ converges.*

In practice, the finite accuracy used in computer calculations seems to prevent convergence to a saddle point.

4.4.6 Remark on complexity

The new algorithm can be restated only in terms of the empirical covariance matrix S defined in (4.4). For example in (4.24), $Y_1\hat{Z}_1' = S_{13} - \hat{\beta}_{32.4}S_{12} - \hat{\beta}_{34.2}S_{14}$ and the remaining quantities can be expressed similarly. Thus, the sample size does not affect the complexity of the algorithm. The complexity of one of the algorithm's pseudo-variable regression steps is dominated by the solution of the systems of $\mathrm{nsp}(i)$ and $\mathrm{sp}(i)$ linear equations in (4.30) and (4.35), respectively.

4.5 Example data

Table 4.1 presents data on $p = 4$ variables measured on $n = 39$ patients; see Cox and Wermuth (1993, Table 7) and Kauermann (1996, Table 1). If we index the variables in this data set by $W = 1$, $V = 2$, $X = 3$, and $Y = 4$ then the covariance graph model fitted by Kauermann is the one illustrated in Figure 4.1(a). We use the observed marginal correlations and standard deviations to reconstruct the empirical covariance matrix. Then we fit the

Table 4.1: Observed marginal correlations and standard deviations.

	W	V	X	Y
V	0.060			
X	−0.460	0.042		
Y	−0.071	−0.404	−0.334	
SD	5.72	92.00	7.86	2.07

model from Figure 4.1(a) by our new algorithm for ML estimation. Table 4.2 shows that the ML estimates and Kauermann's dual estimates are very similar in this example.

Table 4.2: Marginal correlations and standard deviations from ML (lower half & 6th row) and Kauermann's dual estimation (upper half & 5th row).

ML\dual	W	V	X	Y
W		0	−0.479	0
V	0		0	−0.373
X	−0.475	0		−0.351
Y	0	−0.378	−0.342	
SD_{dual}	5.70	91.6	7.92	2.04
SD_{ML}	5.72	92.0	7.93	2.05

4.6 Conclusion/extensions

The new Iterative Conditional Fitting (ICF) algorithm finds ML estimates in covariance graph models using only standard least squares tools. Thus its implementation is straightforward. Moreover, with probability one, the new algorithm converges to a saddle point or local maximum of the likelihood.

Besides the modification discussed in §4.4.2, another modification which potentially

speeds up ICF consists in using multivariate regressions instead of the univariate regressions $(Y_i \mid Y_{-i})$. The multivariate regressions would be of the form $(Y_C \mid Y_{V\setminus C})$ for some subset $C \subseteq V$. If the subset C is complete with respect to G, i.e. every pair of vertices in C is adjacent, then the conditional distribution $(Y_C \mid Y_{V\setminus C})$ has the form of seemingly unrelated regressions. ML estimation in seemingly unrelated regressions itself generally requires iterative algorithms but if the current estimate of Σ is used as a starting value then a single step in such an algorithm would be sufficient for the extension of our algorithm. Following this idea, one could perform edge-wise updates, i.e. $|C| = 2$, but it might possibly be better to perform updates for cliques C.

Importantly, ICF exploits regression techniques, and not directly the likelihood equations. Hence, it may lend itself to generalization; for example, to the case of a graphical model for marginal independence in which the variables are discrete.

The ICF algorithm has similarities with the Iterative Conditional Modes (ICM) algorithm of Besag (1986). However, ICM obtains maximum *a posteriori* estimates in a Bayesian framework, whereas our ICF maximizes a likelihood function. The difference in the update steps is that in the updates of ICM conditional density functions are maximized, whereas in the updates of ICF one maximizes conditional likelihood functions.

Another related algorithm is the Conditional Iterative Proportional Fitting (CIPF) algorithm of Cramer (1998, 2000). CIPF can be used to fit a model that comprises joint distributions for which a set of conditional distributions are set equal to prescribed conditionals. However, CIPF differs from ICF for covariance graphs because the update steps of ICF do not simply equate a conditional distribution with a prescribed conditional, but rather find a conditional distribution by maximizing a conditional likelihood function. In particular, these maximizers will generally not be the same in two different iterations of ICF.

Chapter 5

ITERATIVE CONDITIONAL FITTING FOR GAUSSIAN ANCESTRAL GRAPH MODELS

Ancestral graph models, introduced by Richardson and Spirtes (2002), generalize both Markov random fields and Bayesian networks. A key feature of ancestral graph models is that the global Markov property is closed under conditioning and marginalization. The conditional independence structures that can be encoded by ancestral graphs coincide with the structures that can arise from a Bayesian network with selection and unobserved variables. Thus, association structures learned via ancestral graph models may be interpreted causally. In this chapter, we consider Gaussian ancestral graph models and present an algorithm for maximum likelihood estimation. We call this new algorithm iterative conditional fitting since as in Chapter 4, each step of the procedure fits a conditional distribution subject to constraints, while a marginal distribution is held fixed.

5.1 Introduction

Markov random fields or equivalently undirected graph models as well as Bayesian networks or equivalently directed acyclic directed graph (DAG) models have found wide-spread application. Well-known generalizations of both undirected graph models and DAG models are the chain graph models, which can be equipped with two alternative Markov properties (Andersson et al., 2001). A different generalization is obtained from ancestral graphs, introduced by Richardson and Spirtes (2002). Whereas chain graphs allow both undirected and directed edges, ancestral graphs have edges of three possible types: undirected and directed edges are complemented by bidirected edges.

In ancestral graphs, m-separation, a natural extension of d-separation, yields a global Markov property that is closed under conditioning and marginalization. Interpreted via this

Markov property, ancestral graphs can encode any conditional independence structures that can arise from a Bayesian network with selection and unobserved variables (Richardson and Spirtes, 2002). Marginalization (forming the marginal distribution of the observed variables) is associated with introducing bi-directed edges; conditioning (on selection variables) is associated with introducing undirected edges. Due to this connection between ancestral graphs and underlying DAGs, ancestral graph models not only generalize undirected graph and DAG models, but also lead to conditional independence structures that could have arisen from DAGs and hence are causally interpretable.

This chapter is a first step towards making ancestral graph methodology available for use in applications. The problem we consider is the problem of estimating or learning the parameters of a given ancestral graph model by maximum likelihood. We restrict ourselves to the Gaussian case, for which Richardson and Spirtes (2002) provided a parameterization, and propose a new estimation algorithm, which extends the algorithm described in Chapter 4. We give this new algorithm the name "iterative conditional fitting" (ICF) since in each step of the procedure, a conditional distribution is estimated, subject to constraints, while a marginal distribution is held fixed. This approach is in duality to the well-known iterative proportional fitting algorithm (Whittaker, 1990, pp. 182–185), in the steps of which a marginal distribution is fitted for a fixed conditional distribution.

The remainder of this chapter is organized as follows. In §5.2 and §5.3 we define ancestral graphs and their global Markov property, and in §5.4 we introduce Gaussian ancestral graph models. In §5.5 we present the ICF algorithm. We implemented the algorithm in a function library for the statistical programming system R. Its use is demonstrated in an example session in §5.6. We conclude in §5.7.

5.2 Ancestral graphs

Consider a graph $G = (V, E)$ with the vertex set V and the edge set E containing three types of edge, *undirected* ($-$), *directed* ($\rightarrow$) and *bidirected* ($\leftrightarrow$). However, the graph G is not allowed to have an edge from a vertex to itself or more than one edge between a given pair of vertices. We use the following terminology to describe relations between two vertices

$i, j \in V$ in G:

$$\text{If} \left\{ \begin{array}{c} i - j \\ i \leftrightarrow j \\ i \rightarrow j \end{array} \right\} \text{ then } i \text{ is a } \left\{ \begin{array}{c} \textit{neighbor} \\ \textit{spouse} \\ \textit{parentchild} \end{array} \right\} \text{ of } j.$$

We denote the set of neighbors of a vertex i as $\mathrm{ne}(i)$, the set of spouses as $\mathrm{sp}(i)$ and the set of parents as $\mathrm{pa}(i)$. For vertex sets $A \subseteq V$, we define $\mathrm{ne}(A) = \cup(\mathrm{ne}(i) \mid i \in A)$ and similarly $\mathrm{sp}(A)$, $\mathrm{pa}(A)$.

A sequence of edges between two vertices i and j in G is an ordered (multi)set of edges $\langle e_1, \ldots, e_m \rangle$, such that there exists a sequence of vertices (not necessarily distinct) $\langle i = i_1, \ldots, i_{m+1} = j \rangle$, where edge e_{i_k} has endpoints i_k, i_{k+1}. A sequence of edges for which the corresponding sequence of vertices contains no repetitions is called a *path*. A path of the form $i \rightarrow \cdots \rightarrow j$, on which every edge is of the form $\rightarrow$, with the arrowheads pointing toward j, is a *directed path from i to j*. A directed path from a vertex i to a vertex j followed by the edge $j \rightarrow i$ is called a (fully) *directed cycle*. A special case of the ancestral graphs defined below are directed acyclic graphs (DAG), in which all edges are directed, and there are no directed cycles.

A vertex i is said to be an *ancestor* of a vertex j, denoted $i \in \mathrm{an}(j)$, if either there is a directed path $i \rightarrow \cdots \rightarrow j$ from i to j, or $i = j$. For a vertex set $A \subseteq V$, we define $\mathrm{an}(A) = \cup(\mathrm{an}(i) \mid i \in A)$.

Definition 5.1 (Richardson and Spirtes, 2002, §3). *A graph $G = (V, E)$ with undirected, directed and bidirected edges is an* ancestral graph *if for all $i \in V$ it holds that*

(1) if $\mathrm{ne}(i) \neq \emptyset$ *then* $\mathrm{pa}(i) \cup \mathrm{sp}(i) = \emptyset$;

(2) $i \notin \mathrm{an}(\mathrm{pa}(i) \cup \mathrm{sp}(i))$.

In words, condition (1) states that if there is an undirected edge with endpoint i then there may not exist a directed or bidirected edge with an arrowhead at i, and condition (2) states that there may not be a directed path from a vertex i to one of its parents or spouses. Condition (2) may be restated equivalently as (1) there are no directed cycles, and (2) no spouses are ancestors.

An example of an ancestral graph with vertex set $V = \{0, 1, 2, 3, 4\}$ is given in Figure 5.1. Additional examples can be found in (Richardson and Spirtes, 2002, e.g. Fig. 3, 6, 7 and 12). Note also that any DAG and any undirected graph is an ancestral graph.

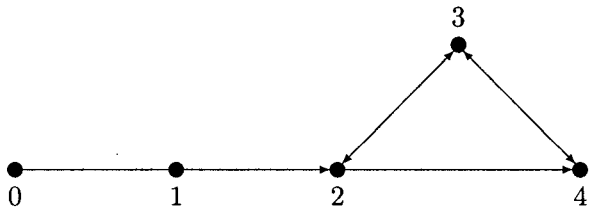


Figure 5.1: An ancestral graph.

Condition (1) implies that an ancestral graph can be decomposed into an undirected part and a part with only directed and bidirected edges (Richardson and Spirtes, 2002, §3.2). Let $\text{un}_G = \{i \in V \mid \text{pa}(i) \cup \text{sp}(i) = \emptyset\}$. Then the subgraph $G_{\text{un}_G} = [\text{un}_G, E \cap (\text{un}_G \times \text{un}_G)]$ induced by un_G is an undirected graph, and any edge between $i \in \text{un}_G$ and $j \notin \text{un}_G$ is directed as $i \to j$. Furthermore, the induced subgraph $G_{V \setminus \text{un}_G}$ contains only directed and bidirected edges. In Figure 5.1, $\text{un}_G = \{0, 1\}$.

5.3 Global Markov property

Pearl's (1988) d-separation criterion for DAGs can be extended to ancestral graphs. A nonendpoint vertex i on a path is a *collider on the path* if the edges preceding and succeeding i on the path have an arrowhead at i, that is, $\to i \leftarrow$, $\to i \leftrightarrow$, $\leftrightarrow i \leftarrow$, $\leftrightarrow i \leftrightarrow$. A nonendpoint vertex i on a path which is not a collider is a *noncollider on the path.* A path between vertices i and j in an ancestral graph G is said to be *m-connecting given a set* C (possibly empty), with $i, j \notin C$, if:

(1) every noncollider on the path is not in C, and

(2) every collider on the path is in $\text{an}(C)$.

If there is no path m-connecting i and j given C, then i and j are said to be m-separated given C. Sets A and B are m-separated given C, if for every pair i, j, with $i \in A$ and $j \in B$, i and j are m-separated given C (A, B, C are disjoint sets; A, B are nonempty). This is an extension of Pearl's d-separation criterion to ancestral graphs in that in a DAG, a path is d-connecting if and only if it is m-connecting.

Let $G = (V, E)$ be an ancestral graph whose vertices V are identified with random variables $(Y_i \mid i \in V)$ and let P be the joint probability distribution of $(Y_i \mid i \in V)$. If $Y_A \perp\!\!\!\perp Y_B \mid Y_C$ whenever A and B are m-separated given C, then P is said to satisfy the *global Markov property for* G, or to be *globally Markov with respect to* G. Here $Y_A = (Y_i \mid i \in A)$ for $A \subseteq V$. For the joint distribution P to be globally Markov with respect to the graph G in Figure 5.1, the conditional independences $Y_0 \perp\!\!\!\perp Y_{234} \mid Y_1$, $Y_1 \perp\!\!\!\perp Y_3$, $Y_1 \perp\!\!\!\perp Y_4 \mid Y_2$ must hold. The global Markov property in addition implies for example $Y_1 \perp\!\!\!\perp Y_3 \mid Y_0$ and $Y_1 \perp\!\!\!\perp Y_4 \mid Y_{02}$ but these are consequences of the previous conditional independences.

In this example, if two vertices i and j are not adjacent, then some conditional independence $Y_i \perp\!\!\!\perp Y_j \mid Y_C$, $C \subseteq V$, holds. Ancestral graphs for which this is true are called *maximal* (Richardson and Spirtes, 2002, §3.7). In fact, by adding bidirected edges, any non-maximal ancestral graph can be converted into a unique maximal ancestral graph without changing the independence model implied by the global Markov property.

The main motivation for considering ancestral graphs is that they can encode conditional independence structures arising from DAGs with selection and unobserved variables. We illustrate this by an example. Consider the DAG in Figure 5.2. Assume that the variables u_{23} and u_{34} are unobserved and that variable s_{01} is a selection variable. If we form the conditional distribution $(0\ 1\ 2\ 3\ 4 \mid s_{01})$, with unobserved variables marginalized out and selection variables conditioned on, then the conditional independences holding in $(0\ 1\ 2\ 3\ 4 \mid s_{01})$ are exactly those implied by the global Markov property of the graph G from Figure 5.1. For details on this connection between DAGs and ancestral graphs see Richardson and Spirtes (2002), who also discuss the relationship between ancestral graphs and related graphical concepts such as the MC-graphs of Koster (1999, 2002) and the summary graphs of Cox and Wermuth (1996).

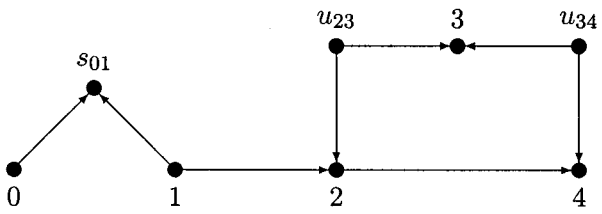


Figure 5.2: DAG with selection variable s_{01} and the unobserved variables u_{23} and u_{34}.

5.4 *Gaussian ancestral graph models*

Suppose now that the variables $(Y_i \mid i \in V)$ jointly follow a centered Gaussian $\equiv$ normal distribution $\mathcal{N}_V(0, \Sigma) \in \mathbb{R}^V$, where $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{V \times V}$ is the unknown positive definite covariance matrix. Let $\mathbf{P}(V)$ be the cone of all positive definite $V \times V$ matrices and let $\mathbf{P}(G)$ be the subcone of all matrices $\Sigma \in \mathbf{P}(V)$ such that $\mathcal{N}_V(0, \Sigma)$ is globally Markov with respect to the given ancestral graph G. The *Gaussian ancestral graph model* based on G is the family of all normal distributions

$$\mathbf{N}(G) = \{\mathcal{N}_V(0, \Sigma) \mid \Sigma \in \mathbf{P}(G)\}. \tag{5.1}$$

As shown in Richardson and Spirtes (2002, §8.4), the model $\mathbf{N}(G)$ forms a curved exponential family.

5.4.1 *Parameterization*

Richardson and Spirtes (2002, §8) provide a parameterization of the Gaussian ancestral graph model $\mathbf{N}(G)$. This parameterization associates one parameter with each vertex in V and each edge in E. Let $\Lambda = (\lambda_{ij})$ be a positive definite $\mathrm{un}_G \times \mathrm{un}_G$ matrix such that $\lambda_{ij} \neq 0$ only if $i = j$ or $i - j$. Recall that $i, j \in \mathrm{un}_G$ can only be adjacent by an undirected edge. Let $\Omega = (\omega_{ij})$ be a positive definite $(V \setminus \mathrm{un}_G) \times (V \setminus \mathrm{un}_G)$ matrix such that $\omega_{ij} \neq 0$ only if $i = j$ or $i \leftrightarrow j$. Finally, let $B = (\beta_{ij})$ be a $V \times V$ matrix such that $\beta_{ij} \neq 0$ only if $j \to i$. Note that the $\mathrm{un}_G \times V$ submatrix of B must be zero, i.e. $B_{\mathrm{un}_G, V} = 0$, because no vertex in un_G

has a parent. With the parameter matrices Λ, B, Ω, we can define the covariance matrix

$$\Sigma = (I_V - B)^{-1} \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & \Omega \end{pmatrix} \left((I_V - B)^{-1}\right)', \tag{5.2}$$

which satisfies $\Sigma \in \mathbf{P}(G)$. Equivalently said, the normal distribution $\mathcal{N}_V(0, \Sigma)$ is globally Markov with respect to the considered ancestral graph G. If G is maximal, then for any $\Sigma \in \mathbf{P}(G)$ there exist unique Λ, Ω, B of the above type such that (5.2) holds.

The population interpretation of the parameters is the following: First, the parameter matrix Λ clearly forms an inverse covariance matrix for the undirected graph G_{un_G}. Second, just as for Gaussian DAG models, the parameter β_{ij}, associated with a directed edge $j \to i$ is the regression coefficient for variable j in the regression of variable i on its parents $\text{pa}(i)$. Third, the parameter ω_{ii} is the conditional variance of the conditional distribution $(Y_i \mid Y_{\text{pa}(i)})$, or equivalently $\omega_{ii} = \text{Var}[\varepsilon_i]$, where

$$\varepsilon_i = Y_i - \sum_{j \in \text{pa}(i)} \beta_{ij} Y_j. \tag{5.3}$$

The parameter ω_{ij} for $i \leftrightarrow j$ is the covariance between the residuals ε_i and ε_j. We illustrate the parameterization by showing in Figure 5.3 the parameters for the graph from Figure 5.1.

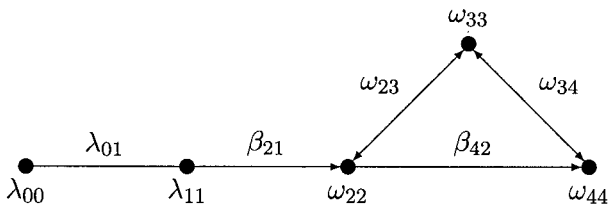


Figure 5.3: Parameters of a Gaussian ancestral graph model.

5.4.2 Maximum likelihood estimation

This chapter considers the estimation of the unknown parameter Σ, or equivalently Λ, Ω, B, of a Gaussian ancestral graph model $\mathbf{N}(G)$ based on a sample of i.i.d. observations from

$\mathbf{N}(G)$. We assume that G is maximal. Let the i.i.d. copies in the sample be indexed by the set N, which can be interpreted as indexing the subjects on which we observed the variables in V. Then we can group the observed random vectors in the sample as columns in the $V \times N$ matrix Y, which means that Y_{im} represents the observation of the i-th variable on the m-th subject. Finally, the sample size is $n = |N|$ and the number of variables is $p = |V|$.

Since our model assumes a zero mean, the empirical covariance matrix is defined to be

$$S = \frac{1}{n} YY' \in \mathbb{R}^{V \times V}. \tag{5.4}$$

We shall assume that $n \geq p$ such that S is positive definite with probability one. Note that the case where the model also includes an unknown mean vector $\mu \in \mathbb{R}^V$ can be treated by estimating μ by the empirical mean vector $\bar{Y} \in \mathbb{R}^V$, i.e. the vector of the row means of Y. The empirical covariance matrix would then be the matrix

$$\tilde{S} = \frac{1}{n}(Y - \bar{Y} \otimes 1_N)(Y - \bar{Y} \otimes 1_N)' \in \mathbb{R}^{V \times V}, \tag{5.5}$$

where $1_N = (1, \ldots, 1) \in \mathbb{R}^N$ and $\otimes$ is the Kronecker product. Estimation of Σ would proceed in the same way as in the centered case by using $\tilde{S}$ instead of S; the only change being that $n \geq p + 1$ ensures almost sure positive definiteness of $\tilde{S}$.

The density function of Y with respect to the Lebesgue measure is the function $f_\Sigma : \mathbb{R}^{V \times N} \to \mathbb{R}$, which can be expressed as

$$\begin{aligned} f_\Sigma(y) &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1} yy')\} \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\frac{n}{2}\mathrm{tr}(\Sigma^{-1} S)\}, \end{aligned} \tag{5.6}$$

see e.g. Edwards (2000, §3.1). Considered as a function of the unknown parameters for fixed data y this gives the likelihood function of the Gaussian ancestral graph model $\mathbf{N}(G)$ as the mapping $L : \mathbf{P}(G) \to \mathbb{R}$ where $L(\Sigma) = f_\Sigma(y)$. In ML estimation, the parameter Σ is estimated by the maximizer $\hat{\Sigma}$ of the likelihood L. Usually, one considers more conveniently the maximization of the log-likelihood $\ell = \log L$, which, ignoring an additive constant, takes the form

$$\ell(\Sigma) = -\frac{n}{2} \log |\Sigma| - \frac{n}{2} \mathrm{tr}\{\Sigma^{-1} S\}. \tag{5.7}$$

Positive definiteness of S guarantees the existence of the global maximum of $\ell(\Sigma)$ over $\mathbf{P}(G)$ but there may be multiple local maxima (compare Chapter 2).

5.4.3 *Employing the decomposition of an ancestral graph*

As described in Richardson and Spirtes (2002, §8.5), the decomposition of an ancestral graph G into an undirected and a directed-bidirected part is accompanied by a factorization of the density function of a distribution in the associated model $\mathbf{N}(G)$. More precisely, if $\Sigma \in \mathbf{P}(G)$, then

$$f_\Sigma(y) = f_\Lambda(y_{\mathrm{un}_G}) f_{B,\Omega}(y_{V\setminus \mathrm{un}_G} \mid y_{\mathrm{un}_G}). \tag{5.8}$$

Here $f_\Lambda(y_{\mathrm{un}_G})$ is the marginal density of Y_{un_G}, and $f_{B,\Omega}(y_{V\setminus \mathrm{un}_G} \mid y_{\mathrm{un}_G})$ is the conditional density of $(Y_{V\setminus \mathrm{un}_G} \mid Y_{\mathrm{un}_G})$. From (5.8) and variation independence of Λ and (B,Ω) it follows that we can find the MLE of Λ by maximizing the marginal likelihood function $L(\Lambda) = f_\Lambda(y_{\mathrm{un}_G})$. Since $Y_{\mathrm{un}_G} \sim \mathcal{N}(0,\Lambda)$, this is precisely fitting an undirected graph model based on the graph G_{un_G} using only the observations for variables in un_G, that is $Y_{\mathrm{un}_G,N}$. Thus, the MLE $\hat{\Lambda}$ of Λ can be obtained by iterative proportional fitting, as described for example in Whittaker (1990, pp. 182–185).

In order to find the MLE $(\hat{B},\hat{\Omega})$ of (B,Ω) we can maximize the conditional likelihood function

$$L(B,\Omega) = f_{B,\Omega}(y_{V\setminus \mathrm{un}_G} \mid y_{\mathrm{un}_G}). \tag{5.9}$$

It is easy to see that the global Markov property for the graph G implies that

$$L(B,\Omega) = f_{B,\Omega}(y_{V\setminus \mathrm{un}_G} \mid y_{\mathrm{pa}(V\setminus \mathrm{un}_G)\cap \mathrm{un}_G}). \tag{5.10}$$

Thus, only a subset $\mathrm{db}_G = [V \setminus \mathrm{un}_G] \cup \mathrm{pa}(V \setminus \mathrm{un}_G)$ of the variables is needed for estimating (B,Ω); i.e. we are using the observations $Y_{\mathrm{db}_G,N}$. The set db_G is the set of all vertices i in G that are the endpoint of at least one directed or bidirected edge, i.e. there is an edge $i \to j$, $i \leftarrow j$ or $i \leftrightarrow j$. For the graph G from Figure 5.1, the induced subgraph G_{db_G} is shown in Figure 5.4.

In the next section, we present an algorithm for estimating (B,Ω). This algorithm extends the idea developed in Chapter 4 of this thesis, which in fact considers ancestral graphs with bidirected edges only. The idea is to iteratively fit a conditional distribution for a fixed marginal distribution. Thus we call this algorithm *iterative conditional fitting.*

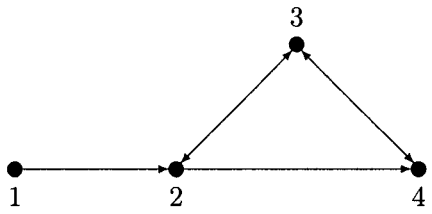


Figure 5.4: The graph $G_{\mathrm{db}G}$.

Note the duality with the iterative proportional fitting algorithm, in which, cycling through a list of cliques, the marginal distribution over a clique $C \subseteq V$ is fitted while fixing the conditional distribution $(Y_{V\setminus C} \mid Y_C)$.

5.5 *Iterative conditional fitting*

5.5.1 The general algorithm

Let G be a maximal ancestral graph. The idea of iterative conditional fitting (ICF) is to repeatedly iterate through all vertices $i \in V$, and

(1) Fix the marginal distribution for $Y_{-i} = Y_{V\setminus\{i\}}$.

(2) Fit the conditional distribution $(Y_i \mid Y_{-i})$ under the constraints implied by the Gaussian ancestral graph model $\mathbf{N}(G)$.

(3) Find a new estimate of Σ from the estimated parameters of the conditional distribution $(Y_i \mid Y_{-i})$ and the fixed parameters of the marginal distribution of Y_{-i}.

In Chapter 4, where the graph G only contained bidirected edges, the problem of fitting $(Y_i \mid Y_{-i})$ under constraints could be rephrased as a least squares regression problem. Here, however, where G also contains directed edges, the consideration of $(Y_i \mid Y_{-i})$ is complicated. Fortunately, we can "remove" directed edges by forming residuals as in (5.3), and considering the conditional distribution $(Y_i \mid \varepsilon_{-i})$ as presented in the following. Such a residual trick can already be found in Telser (1964).

Before formulating ICF for ancestral graphs, we note two facts. First, the maximization of $L(B, \Omega)$ from (5.9) is by (5.8) equivalent to maximizing $L(\Sigma) = L(\Lambda, B, \Omega)$ with holding Λ fixed to some feasible value Λ_0, which could be, for example, the identity matrix or the matrix found by iterative proportional fitting as described in §5.4.3. Second, fixing $\Lambda = \Lambda_0$, the matrix $\varepsilon = (I_V - B)Y$ has i-th row equal to the residual ε_i defined in (5.3) and each column of ε has covariance matrix $\Psi = (\psi_{ij})$ equal to

$$\Psi = \begin{pmatrix} \Lambda_0^{-1} & 0 \\ 0 & \Omega \end{pmatrix}. \tag{5.11}$$

From this and the fact $B_{\mathrm{un}_G, V} = 0$ we see that, in order to estimate (B, Ω) we need only cycle through the vertices $i \notin \mathrm{un}_G$.

Next we compute the conditional distribution $(Y_i \mid \varepsilon_{-i})$ for $i \notin \mathrm{un}_G$. This distribution is obviously Gaussian and its conditional variance equals

$$\mathrm{Var}[Y_i \mid \varepsilon_{-i}] = \omega_{ii.-i}, \tag{5.12}$$

which is defined as

$$\omega_{ii.-i} = \omega_{ii} - \Omega_{i,-i}(\Omega_{-i,-i})^{-1}\Omega_{-i,i}. \tag{5.13}$$

Equation (5.12) holds because, for $i \notin \mathrm{un}_G$

$$\begin{aligned} \mathrm{Var}[Y_i \mid \varepsilon_{-i}] &= \mathrm{Var}[\varepsilon_i \mid \varepsilon_{-i}] \\ &= \psi_{ii} - \Psi_{i,-i}(\Psi_{-i,-i})^{-1}\Psi_{-i,i} \\ &= \omega_{ii} - \Omega_{i,-i}(\Omega_{-i,-i})^{-1}\Omega_{-i,i}. \end{aligned} \tag{5.14}$$

Note that when writing $\Omega_{-i,-i}$ we mean the $[V \setminus (\mathrm{un}_G \cup \{i\})] \times [V \setminus (\mathrm{un}_G \cup \{i\})]$ submatrix of Ω. The normal distribution $(Y_i \mid \varepsilon_{-i})$ is now specified by (5.12) and the conditional expectation, which equals

$$\begin{aligned} \mathrm{E}[Y_i \mid \varepsilon_{-i}] &= \sum_{j \in \mathrm{pa}(i)} \beta_{ij} \mathrm{E}[Y_j \mid \varepsilon_{-i}] + \mathrm{E}[\varepsilon_i \mid \varepsilon_{-i}] \\ &= \sum_{j \in \mathrm{pa}(i)} \beta_{ij} Y_j + \sum_{k \in \mathrm{sp}(i)} \omega_{ik} Z_k, \end{aligned} \tag{5.15}$$

where the *pseudo-variable* Z_k is equal to the k-th row in

$$Z_{\mathrm{sp}(i)} = [(\Omega_{-i,-i})^{-1}]_{\mathrm{sp}(i),-i}\, \varepsilon_{-i}. \tag{5.16}$$

The derivation of (5.15) relies on two facts. First, for any $j \in \mathrm{pa}(i)$, Y_j is a function of $\varepsilon_{\mathrm{an}(i)\setminus\{i\}}$ and thus, $\mathrm{E}[Y_j \mid \varepsilon_{-i}] = Y_j$. Second, for $i \notin \mathrm{un}_G$ the residual covariance $\psi_{ij} = 0$ if $i \not\to j$. Thus, $\Psi_{i,\mathrm{nsp}(i)} = 0$, which implies that for $i \notin \mathrm{un}_G$:

$$\begin{aligned} \mathrm{E}[\varepsilon_i \mid \varepsilon_{-i}] &= \Psi_{i,-i}(\Psi_{-i,-i})^{-1}\, \varepsilon_{-i} \\ &= \Psi_{i,\mathrm{sp}(i)}[(\Psi_{-i,-i})^{-1}]_{\mathrm{sp}(i),-i}\, \varepsilon_{-i} \\ &= \Omega_{i,\mathrm{sp}(i)}[(\Omega_{-i,-i})^{-1}]_{\mathrm{sp}(i),-i}\, \varepsilon_{-i} \\ &= \sum_{k \in \mathrm{sp}(i)} \omega_{ik} Z_k. \end{aligned} \tag{5.17}$$

After this preparation we are now ready to formulate ICF for ancestral graphs to find the MLE $(\hat{B}, \hat{\Omega})$. Until convergence, for each $i \notin \mathrm{un}_G$:

1. Fix $\Omega_{-i,-i}$ and all $B_{j,\mathrm{pa}(j)} = (\beta_{j\ell} \mid \ell \in \mathrm{pa}(j))$ for $j \neq i$;

2. Use the fixed $\beta_{j\ell}$ to compute the residuals ε_j for $j \neq i$ from (5.3);

3. Use the fixed $\Omega_{-i,-i}$ to compute the pseudo-variables Z_k for $k \in \mathrm{sp}(i)$;

4. Carry out a least squares regression with response variable Y_i and covariates Y_j, $j \in \mathrm{pa}(i)$, and Z_k, $k \in \mathrm{sp}(i)$ to obtain estimates of β_{ij}, $j \in \mathrm{pa}(i)$, ω_{ik}, $k \in \mathrm{sp}(i)$, and $\omega_{ii.-i}$;

5. Compute an estimate of ω_{ii} using the new estimates and the fixed parameters from the relation $\omega_{ii} = \omega_{ii.-i} + \Omega_{i,-i}(\Omega_{-i,-i})^{-1}\Omega_{-i,i}$, compare (5.13).

After steps (1)-(5), we move on to the next vertex in $V \setminus \mathrm{un}_G$. The procedure is continued until convergence.

5.5.2 *Convergence*

It is easy to see that this ICF algorithm is an iterative partial maximization algorithm (compare Appendix A) since in the i-th step we maximize the conditional likelihood $L(B, \Omega)$ from (5.9) over the section in the parameter space defined by fixing the parameters $\Omega_{-i,-i}$, and $B_{j,\mathrm{pa}(j)}$, $j \neq i$. The same reasoning as in Chapter 4, §4.4.5 applies and we can deduce that for any feasible starting value the algorithm produces a sequence of estimates, for which each accumulation point is a local maximum or a saddle point of the likelihood. Furthermore, evaluating the likelihood at each accumulation point must give the same value.

5.5.3 *Applying ICF to DAGs*

It is well known that the MLE of the parameters of a Gaussian DAG model can be found by carrying out a finite number of regressions (see e.g. Wermuth (1980), Goldberger (1991), or Andersson and Perlman (1998)). DAG models form a special case of ancestral graph models so we can also apply ICF to a Gaussian DAG model. If the graph G is a DAG then $\mathrm{sp}(i) = \emptyset$ for all $i \in V$. Therefore, the conditional distribution of $(Y_i \mid \varepsilon_{-i})$ is fitted by regressing solely on the parents Y_j, $j \in \mathrm{pa}(i)$; compare (5.15). Thus the least squares regression carried out in the i-th step of ICF is always the same $(Y_i \mid Y_{\mathrm{pa}(i)})$ since it involves no pseudo-variables, which could change from one iteration to the other. This means that ICF reduces to the standard approach of fitting Gaussian DAG models if the ancestral graph under consideration is in fact a DAG.

5.5.4 *ICF in an example*

We illustrate estimation of (B, Ω) by ICF in the example of the ancestral graph depicted in Figure 5.1. The set un_G equals $\{0, 1\}$ and in fact only the variables $\mathrm{db}_G = \{1, 2, 3, 4\}$ are relevant for estimating (B, Ω), compare Figure 5.4. The iteration steps, described in items (1)-(5) in §5.5.1, have to be carried only for $i \in V \setminus \mathrm{un}_G = \{2, 3, 4\}$. In Figure 5.5, we show the response variable Y_i in the i-th ICF update step as a filled circle. The remaining variables are depicted as unfilled circles. A vertex labelled by a number j represents the

random variable Y_j and a vertex labelled by Z_j represents the pseudo-variable defined in (5.16). Finally, we use directed edges pointing from a covariate Y_j or Z_j to the response Y_i to indicate the structure of the least squares regression that has to be performed in the i-th ICF step. Note that we suppress the irrelevant variable 0.

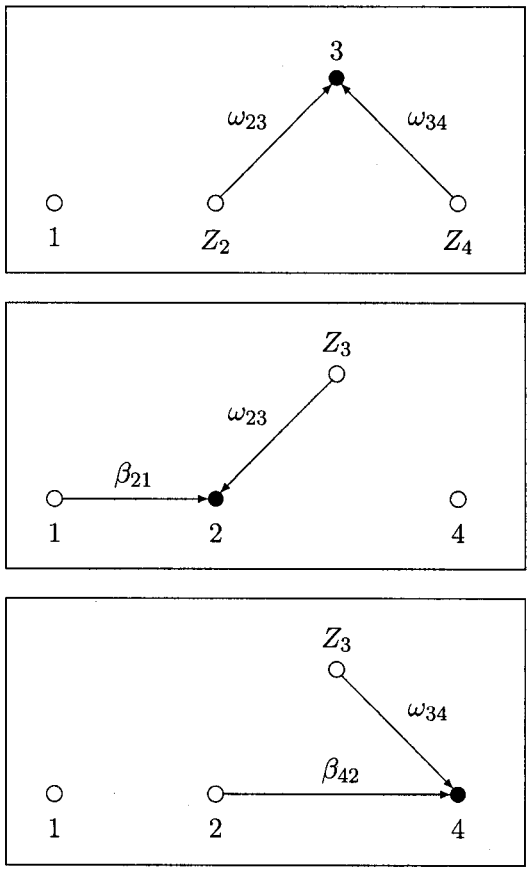


Figure 5.5: Illustration of the ICF update steps.

5.6 An implementation

The statistical programming language R (Ihaka and Gentleman, 1996) provides a freeware environment for programming in interpreted code building on a large number of available routines. The team of developers of R provided a framework for writing extension libraries

for R. As part of the "graphical models in R" initiative (Lauritzen, 2002), Marchetti and Drton developed a function library called 'ggm', which implements functions for fitting Gaussian graphical models and, in particular, provides an implementation of ICF. The package can be downloaded from `http://cran.r-project.org/`.

In the following, we show an example session in R using 'ggm' and data from the R function library 'SIN', which implements the model selection method described in Drton and Perlman (2004). We begin by loading 'ggm', 'SIN', and a data set on noctuid moth trappings. The data comprise $n = 72$ measurements for each one of six variables: the (log-transformed) number of moths caught in a light trap in one night (moth), the minimum night temperature (min), the previous day's maximum temperature (max), the average wind speed during night (wind), the amount of rain during night (rain), and the percentage of starlight obscured by clouds (cloud); compare for example Whittaker (1990, §10.3). We display the correlation matrix and the sample size for these data.

```
> library(ggm)
> library(SIN)
> data(moth)
> moth$corr
        min   max  wind  rain cloud  moth
min    1.00  0.40  0.37  0.18 -0.46  0.29
max    0.40  1.00  0.02 -0.09  0.02  0.22
wind   0.37  0.02  1.00  0.05 -0.13 -0.24
rain   0.18 -0.09  0.05  1.00 -0.47  0.11
cloud -0.46  0.02 -0.13 -0.47  1.00 -0.37
moth   0.29  0.22 -0.24  0.11 -0.37  1.00
> moth$n
[1] 72
```

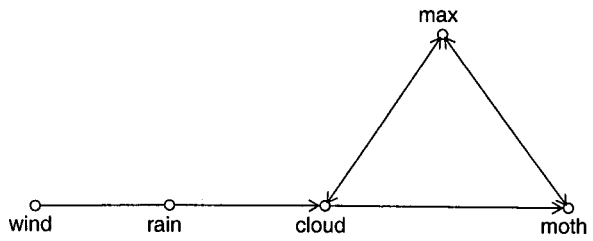
We concentrate on the five variables max, wind, rain, cloud, and moth to fit the model induced by the (maximal) ancestral graph from Figure 5.1. In order to do this we first define the ancestral graph by defining complete subsets for the undirected and bidirected subgraphs, and regression structures for the directed subgraph. The resulting graph is

represented by an adjacency matrix $A = (a_{ij})$ where $a_{ij} = a_{ji} = 1$ if $i - j$, $a_{ij} = a_{ji} = 2$ if $i \leftrightarrow j$, and $a_{ij} = 1$, $a_{ji} = 0$ if $i \rightarrow j$.

```
> mag <- makeAG(ug=UG(~wind*rain),
+               dag=DAG(cloud~rain, moth~cloud),
+               bg=UG(~max*cloud+max*moth))
> mag
      max cloud moth wind rain
max     0     2    2    0    0
cloud   2     0    1    0    0
moth    2     0    0    0    0
wind    0     0    0    0    1
rain    0     1    0    1    0
```

The package 'ggm' also provides a rudimentary tool for plotting graphs.

```
> drawGraph(mag)
```



Now we are able to fit the Gaussian ancestral graph model to the data using the function 'fitAncestralGraph', which returns $\hat{\Sigma}$ as `Shat`, $\hat{\Lambda}$ as `Lhat`, $I_V - \hat{B}$ as `Bhat` and $\hat{\Omega}$ as `Ohat`. The output also includes the deviance statistic `dev`, the degrees of freedom `df` and the number of iterations `it`, that is, the number of full cycles through all $i \notin \mathrm{un}_G$ during ICF. The deviance statistic is computed as

$$\mathtt{dev} = 2\ell(S) - 2\ell(\hat{\Sigma}). \tag{5.18}$$

Note, however, that the ICF estimate $\hat{\Sigma}$ may only be a local maximum (compare Chapter 2).

```
> icf <- fitAncestralGraph(mag, moth$corr, moth$n)
> lapply( icf , round, 2 )
$Shat
          max  wind  rain cloud  moth
  max    1.00  0.00  0.00 -0.02  0.23
  wind   0.00  1.00  0.05 -0.02  0.01
  rain   0.00  0.05  1.00 -0.47  0.18
  cloud -0.02 -0.02 -0.47  1.00 -0.38
  moth   0.23  0.01  0.18 -0.38  1.01

$Lhat
      max wind rain cloud moth
max     0 0.00 0.00     0    0
wind    0 1.00 0.05     0    0
rain    0 0.05 1.00     0    0
cloud   0 0.00 0.00     0    0
moth    0 0.00 0.00     0    0

$Bhat
      max wind rain cloud moth
max     1    0 0.00  0.00    0
wind    0    1 0.00  0.00    0
rain    0    0 1.00  0.00    0
cloud   0    0 0.47  1.00    0
moth    0    0 0.00  0.38    1

$Ohat
        max wind rain cloud moth
max    1.00    0    0 -0.02 0.23
wind   0.00    0    0  0.00 0.00
rain   0.00    0    0  0.00 0.00
cloud -0.02    0    0  0.78 0.00
moth   0.23    0    0  0.00 0.86
```

```
$dev
[1] 10.22
$df
[1] 5
$it
[1] 6
```

Comparing the deviance and the degrees of freedom using the asymptotic distribution of the deviance as $\chi^2_{\mathtt{df}}$ yields a p-value of 0.07 suggesting that the model is not inappropriate. However, the inclusion of the additional directed edge wind $\rightarrow$ moth greatly improves the fit with a deviance of 2.01 over 4 degrees of freedom and an associated p-value of 0.73.

5.7 Conclusion

We have presented ICF = iterative conditional fitting, which is an iterative partial maximization algorithm for fitting Gaussian ancestral graph models. Fitting conditional distributions while fixing marginal distributions, ICF stands in duality with the iterative proportional fitting algorithm, in which marginal distributions are fitted while conditional distributions are fixed. ICF is particularly attractive since if the ancestral graph under consideration is in fact a DAG, then the likelihood is maximized in a finite number of steps performing exactly the regressions commonly used for fitting Gaussian DAG models.

A topic of future work will be using Markov equivalence of ancestral graphs for improving efficiency. As it is true for DAGs, different ancestral graphs may induce the same statistical model, in which case the graphs are called Markov equivalent. Since the update steps of the ICF algorithm depend on the graph itself, it is important to work out which graph in a whole class of Markov equivalent graphs allows for the most efficient fitting of the associated model (see also Chapter 4, §4.2.4).

Finally, ICF has the nice feature that its main idea of decomposing the complicated overall maximization problem into a sequence of simpler optimization problems seems also promising for the development of fitting methodology in the case of discrete variables. This extension will be the subject of future work.

Chapter 6

FUTURE WORK

This thesis solves the problem of maximum likelihood estimation in Gaussian AMP chain graph and Gaussian ancestral graph models. Here, "solving" stands for proposing iterative algorithms for maximizing the respective likelihood function such that convergence guarantees can be given.

The results of this thesis suggest the following future work. First, it should be investigated whether ideas such as multivariate updates in iterative conditional fitting can lead to more efficient algorithms for maximizing the Gaussian likelihood functions. Second, methodology for constructing confidence intervals should be developed to complement the point estimates that can be obtained from the proposed algorithms. Finally, discrete analogs of the Gaussian models should be studied in order to make AMP chain graph and ancestral graph modelling more widely applicable.

BIBLIOGRAPHY

Andersen, H. H., Højbjerre, M., Sørensen, D. and Eriksen, P. S. (1995). *Linear and Graphical Models for the Multivariate Complex Normal Distribution*, vol. 101 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*. Academic Press, New York, 55–66.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In *Essays in Probability and Statistics*. Univ. of North Carolina Press, Chapel Hill, N.C., 1–24.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1** 135–141.

Anderson, T. W. and Olkin, I. (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra Appl.* **70** 147–171.

Andersson, S. A., Madigan, D. and Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.* **28** 33–86.

Andersson, S. A. and Perlman, M. D. (1994). Normal linear models with lattice conditional independence restrictions. In *Multivariate Analysis and its Applications*, vol. 24. Inst. Math. Statist., Hayward, CA.

Andersson, S. A. and Perlman, M. D. (1998). Normal linear regression models with recursive graphical Markov structure. *J. Multivariate Anal.* **66** 133–187.

Banerjee, M. and Richardson, T. S. (2003). On a dualization of graphical Gaussian models: A correction note. *Scand. J. Statist.* **30** 817–820.

Besag, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

Binkley, J. K. and Nelson, C. H. (1988). A note on the efficiency of Seemingly Unrelated Regression. *American Statistician* **42** 137–139.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

Boyles, R. A. (1983). On the convergence of the EM algorithm. *J. Roy. Statist. Soc. B* **45** 47–50.

Capitanio, A., Azzalini, A. and Stanghellini, E. (2003). Graphical models for skew-normal variates. *Scand. J. Statist.* **30** 129–144.

Caputo, A., Foraita, R., Klasen, S. and Pigeot, I. (2003). Undernutrition in Benin - An analysis based on graphical models. *Social Science & Medicine* **56** 1677–1691.

Caputo, A., Heinicke, A. and Pigeot, I. (1999). A graphical chain model derived from a model selection strategy for the sociologists graduates study. *Biom. J.* **41** 217–234.

Corradi, F., Lago, G. and Stefanini, F. M. (2003). The evaluation of DNA evidence in pedigrees requiring population inference. *J. R. Statist. Soc. A* **166** 425–440.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* Statistics for Engineering and Information Science, Springer-Verlag, New York.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Stat. Sci.* **8** 204–218,247–277.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation.* Chapman and Hall, London.

Cramer, E. (1998). Conditional iterative proportional fitting for Gaussian distributions. *J. Multivariate Anal.* **65** 261–276.

Cramer, E. (2000). Probability measures with given marginals and conditionals: *I*-projections and conditional iterative proportional fitting. *Statist. Decisions* **18** 311–329.

Cramér, H. (1999). *Mathematical Methods of Statistics. Reprint of the 1946 original.* Princeton University Press, Princeton, NJ.

Dhrymes, P. J. (1970). *Econometrics: Statistical Foundations and Applications.* Harper & Row, New York.

Didelez, V., Pigeot, I., Dean, K. and Wister, A. (2002). A comparative analysis of graphical interaction and logistic regression modelling: self-care and coping with a chronic illness in later life. *Biom. J.* **44** 410–432.

Drton, M. and Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, to appear.

Edwards, C. H. (1994). *Advanced Calculus of Several Variables.* Dover Publications, New York.

Edwards, D. M. (2000). *Introduction to Graphical Modelling*, 2nd ed. Springer-Verlag, New York.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7** 601–620.

Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Statist.* **17** 333–353.

Goldberger, A. (1991). *A Course in Econometrics.* Harvard University Press, Cambridge, Massachusetts.

Green, P. J., Hjort, N. L. and Richardson, S. (2003). *Highly structured stochastic systems.* Oxford University Press, Oxford, UK.

Greene, W. H. (1997). *Econometric Analysis*, 3rd ed. Prentice Hall, Upper Saddle River, New Jersey.

Hartemink, A., Gifford, D., Jaakkola, T. and Young, R. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Pacific Symposium on Biocomputing 2001 (PSB01)* (R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale and T. Klein, eds.). World Scientific, New Jersey, 422–433.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective.* Springer-Verlag, New York.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5** 299–314.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, to appear.

Kauermann, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.* **23** 105–116.

Koster, J. T. A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scand. J. Statist.* **26** 413–431.

Koster, J. T. A. (2002). Marginalizing and conditioning in graphical models. *Bernoulli* **8** 817–840.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford Statistical Science Series, Clarendon Press, Oxford, UK.

Lauritzen, S. L. (2001). Causal inference from graphical models. In *Complex stochastic systems (Eindhoven, 1999)*, vol. 87 of *Monogr. Statist. Appl. Probab.* Chapman & Hall/CRC, Boca Raton, FL, 63–107.

Lauritzen, S. L. (2002). Graphical models in R: A new initiative within the R project. *R News* **2** 39. `http://cran.r-project.org/doc/Rnews/`.

Lauritzen, S. L. and Richardson, T. S. (2002). Chain graph models and their causal interpretations. *J. R. Statist. Soc. B* **64** 321–361.

Lauritzen, S. L. and Sheehan, N. A. (2003). Graphical models for genetic analyses. *Statist. Sci.* **18** 489–514.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17** 31–57.

Levitz, M., Perlman, M. D. and Madigan, D. (2001). Separation and completeness properties for AMP chain graph Markov models. *Ann. Statist.* **29** 1751–1784.

Levitz, M., Perlman, M. D. and Madigan, D. (2003). Correction: "Separation and completeness properties for AMP chain graph Markov models" [Ann. Statist. **29** (2001), 1751–1784]. *Ann. Statist.* **31** 348.

Magnus, J. R. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance matrix. *J. Econometrics* **7** 281–312.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278.

Meng, X.-L. and Rubin, D. B. (1996). Efficient methods for estimation and testing with seemingly unrelated regressions in the presence of latent variables and missing observations. In *Bayesian analysis in statistics and econometrics.* Wiley, New York, 215–227.

Mohamed, W. N., Diamond, I. and Smith, P. F. (1998). The determinants of infant mortality in Malaysia: a graphical chain modelling approach. *J. R. Statist. Soc. A* **161** 349–366.

Oberhofer, W. and Kmenta, J. (1974). A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica* **42** 579–590.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann.

Pearl, J. (2000). *Causality.* Cambridge University Press, Cambridge, UK.

Pearl, J. and Wermuth, N. (1994). When can association graphs admit a causal interpretation? In *Selecting models from data: Artificial intelligence and Statistics IV* (P. Cheeseman et al., ed.), vol. 89 of *Lecture Notes in Statistics.* Springer-Verlag, New York, 205–214.

Percy, D. F. (1996). Zellner's influence on multivariate linear models. In *Bayesian Analysis in Statistics and Econometrics* (D. A. Berry, K. M. Chaloner and J. K. Geweke, eds.). Wiley, New York, 203–14.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York.

Reeds, J. A. (1985). Asymptotic number of roots of Cauchy location likelihood equations. *Ann. Statist.* **13** 775–784.

Richardson, T. S. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.* **30** 145–157.

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030.

Richardson, T. S. and Spirtes, P. (2003). Causal inference via ancestral graph models. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.). Oxford University Press, Oxford, UK, 83–105.

Rochon, J. (1996a). Accounting for covariates observed post-randomization for discrete and continuous repeated measures data. *J. R. Statist. Soc.* B **58** 205–19.

Rochon, J. (1996b). Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics* **52** 740–50.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14** 138–150.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, prediction, and search*, 2nd ed. Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA. Second edition of 1993, Springer, Lecture Notes in Statistics 81.

Srivastava, V. K. and Giles, D. E. A. (1987). *Seemingly Unrelated Regression Equation Models.* Marcel Dekker, New York.

Sugiura, N. and Gupta, A. K. (1987). Maximum likelihood estimates for Behrens-Fisher problem. *J. Japan Statist. Soc.* **17** 55–60.

Telser, L. G. (1964). Iterative estimation of a set of linear regression equations. *J. Amer. Statist. Assoc.* **59** 845–862.

van Garderen, K. J. (1997). Curved exponential models in econometrics. *Econometric Theory* **13** 771–790.

Verbyla, A. P. and Venables, W. N. (1988). An extension of the growth curve model. *Biometrika* **75** 129–38.

Waddell, P. J. and Kishino, H. (2000). Cluster inferences methods and graphical models evaluated on nci60 microarray gene expression data. *Genome Informatics* **11**.

Wang, J., Myklebost, O. and Hovig, E. (2003). Mgraph: graphical model for microarray data analysis. *Bioinformatics* **19** 2210–2211.

Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75** 963–972.

Wermuth, N. (1992). On block-recursive linear regression equations. *Rebrape* **6** 1–56.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. R. Statist. Soc. B* **52** 21–50, 51–72.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichester.

Wu, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103.

Wu, L. and Perlman, M. D. (2000). Lattice conditional independence models for seemingly unrelated regressions. *Comm. Statist. B—Simulation Comput.* **29** 361–384.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *J. Amer. Stat. Assoc.* **57** 348–368.

Zellner, A. (1963). Estimators for seemingly unrelated regression equations: some exact finite sample results. *J. Amer. Stat. Assoc.* **58** 977–992.

Appendix A

ITERATIVE PARTIAL MAXIMIZATION

All the algorithms for maximum likelihood estimation proposed in this thesis are iterative partial maximization algorithms. Partial maximization refers to a maximization of the likelihood function over a section in the parameter space. In an iterative partial maximization algorithm, one repeatedly performs a sequence of partial maximizations. Such an algorithm is well-defined if each partial maximization problem admits a unique solution. In this appendix, we derive some general convergence results for the class of iterative partial maximization algorithms. In particular, this appendix generalizes Lauritzen (1996, Appendix A.4) by not assuming the existence of a unique local = global maximum of the likelihood function.

Let $\ell : \Theta \to \mathbb{R}$ be a differentiable real-valued function defined over a parameter space $\Theta \subseteq \mathbb{R}^d$. Usually ℓ is a (log-)likelihood function. Let $\theta_0 \in \Theta$ be such that

$$\Theta_0 = \{\theta \in \Theta : \ell(\theta) \geq \ell(\theta_0)\} \tag{A.1}$$

is compact. Thus, as consequence of the continuity of ℓ and the compactness of Θ_0, the global maximum of ℓ over Θ_0 exists.

Let $g_i : \Theta \to \mathbb{R}^{d_i}$, $1 \leq i \leq k$, be mappings defined over the parameter space Θ. For a parameter $\theta^* \in \Theta$, the mappings g_i define sections in the parameter space as

$$\Theta_i(\theta^*) = \{\theta \in \Theta : g_i(\theta) = g_i(\theta^*)\}. \tag{A.2}$$

With each section $\Theta_i(\theta^*)$ is associated a partial maximization

$$\max\{\ell(\theta) : \theta \in \Theta_i(\theta^*)\}. \tag{A.3}$$

We assume that for any $\theta^* \in \Theta$ and $1 \leq i \leq k$ the partial maximization problem (A.3) has a unique solution

$$T_i(\theta^*) = \arg\max\{\ell(\theta) : \theta \in \Theta_i(\theta^*)\}, \tag{A.4}$$

that is,

$$\ell\{T_i(\theta^*)\} > \ell(\theta), \quad \forall \theta \in \Theta_i(\theta^*),\ \theta \neq T_i(\theta^*). \tag{A.5}$$

Furthermore, we assume that the mappings $T_i : \Theta \to \Theta$ are continuous for all $1 \leq i \leq k$. Finally, we suppose that the section-defining functions g_i are such that if $\theta^* = T_i(\theta^*)$ for all $1 \leq i \leq k$, then θ^* solves the likelihood equations, i.e.,

$$\Big(\ \theta^* = T_i(\theta^*),\ \forall i : 1 \leq i \leq k\Big) \quad \Longrightarrow \quad \left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = 0. \tag{A.6}$$

Note that this assumption also implies that the global maximum of ℓ over Θ solves the likelihood equations. In many contexts, as for example in this thesis, it is straightforward to verify (A.6). For general conditions to check (A.6), see Meng and Rubin (1993).

In iterative partial maximization, one iteratively solves all partial maximization problems creating a sequence of parameters as follows. Let $\theta_0 \in \Theta$ be a starting value such that Θ_0 is compact. Define

$$\theta_{n+1} = S(\theta_n) = T_k \cdots T_1(\theta_n), \quad n \geq 0. \tag{A.7}$$

Obviously,

$$\ell(\theta_{n+1}) \geq \ell(\theta_n), \quad \forall n, \tag{A.8}$$

which implies in particular that

$$\theta_n \in \Theta_0, \quad \forall n. \tag{A.9}$$

Lemma A.1. *The sequence* $\big(\ell(\theta_n)\big)_n$ *converges to a limit* $\ell_\infty \in \mathbb{R}$.

Proof. Due to (A.8), the sequence $\big(\ell(\theta_n)\big)_n$ is monotone. The existence of a global maximum of ℓ over Θ_0 implies that the sequence $\big(\ell(\theta_n)\big)_n$ is bounded, thus converges. □

Lemma A.2. *Every accumulation point* θ^* *of the sequence* $(\theta_n)_n$ *is in* Θ_0, *satisfies* $\ell(\theta^*) = \ell_\infty$, *and solves the likelihood equations, i.e.*

$$\left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = 0.$$

Proof. First, since Θ_0 is compact, (A.9) implies that every accumulation point θ^* is in Θ_0. Consider one such θ^* and let $(\theta_{n_r})_r$ be a subsequence of $(\theta_n)_n$ such that

$$\lim_{r\to\infty} \theta_{n_r} = \theta^*. \tag{A.10}$$

By Lemma A.1, every subsequence of $\big(\ell(\theta_n)\big)_n$ converges to ℓ_∞ and due to the continuity of ℓ it holds in particular that

$$\ell(\theta^*) = \lim_{r\to\infty} \ell(\theta_{n_r}) = \lim_{r\to\infty} \ell(\theta_{n_r+1}) = \ell_\infty. \tag{A.11}$$

Using the continuity of ℓ and the partial maximization operators T_i, $1 \le i \le k$, we obtain that

$$\ell\{T_k \cdots T_1(\theta^*)\} = \lim_{r\to\infty} \ell\{T_k \cdots T_1(\theta_{n_r})\} \overset{(A.7)}{=} \lim_{r\to\infty} \ell(\theta_{n_r+1}) \overset{(A.11)}{=} \ell_\infty. \tag{A.12}$$

Moreover, by definition of T_i, $1 \le i \le k$,

$$\ell_\infty = \ell\{T_k \cdots T_1(\theta^*)\} \ge \ell\{T_{k-1} \cdots T_1(\theta^*)\} \ge \ldots \ge \ell\{T_1(\theta^*)\} \ge \ell(\theta^*) = \ell_\infty. \tag{A.13}$$

Therefore, in particular, $\ell\{T_1(\theta^*)\} = \ell(\theta^*)$, which means that θ^* is a solution to the partial maximization problem (A.3) for $i = 1$. Since the maximizer is unique, compare (A.5), it must hold that $T_1(\theta^*) = \theta^*$. By induction over i, it follows that

$$T_i(\theta^*) = \theta^*, \quad \forall 1 \le i \le k, \tag{A.14}$$

which by (A.6) implies that θ^* solves the likelihood equations. □

If the likelihood has only one stationary point, i.e., the likelihood equations only one solution, then under assumption (A.6), this stationary point has to be the global maximizer $\tilde{\theta}$ of ℓ over Θ. In this case, $\tilde{\theta}$ is the only possible accumulation point of $(\theta_n)_n$.

Corollary A.3. *If there exists only one solution to the likelihood equations, then $(\theta_n)_n$ converges to this solution to the unique solution to the likelihood equations, which equals the global maximizer of ℓ over Θ.*

However, the sequence $(\theta_n)_n$ also converges under less restrictive conditions.

Proposition A.4. *Let*

$$\mathcal{A}_\infty = \{\theta^* \in \Theta_0 \,:\, \theta^* \text{ is an accumulation point of } (\theta_n)_n\}\,.$$

If $\mathcal{A}_\infty$ is a finite set, then it is in fact a singleton, $\mathcal{A}_\infty = \{\theta_\infty\}$, and the sequence $(\theta_n)_n$ converges to θ_∞.

Proof. Let

$$B_\varepsilon(\theta) = \{\tilde{\theta} \in \Theta \,:\, \|\tilde{\theta} - \theta\| < \varepsilon\}, \quad \varepsilon > 0, \tag{A.15}$$

be the open ε-ball around θ. Let

$$B_\varepsilon(\mathcal{A}_\infty) = \cup\big(B_\varepsilon(\theta^*) \mid \theta^* \in \mathcal{A}_\infty\big) \tag{A.16}$$

be the union of all the ε-balls around the accumulation points of $(\theta_n)_n$. By definition, $\mathcal{A}_\infty$ contains all accumulation points of $(\theta_n)_n$. Thus, for any $\varepsilon > 0$, the number of elements of the sequence $(\theta_n)_n$ that do not lie in $B_\varepsilon(\mathcal{A}_\infty)$ must be finite, i.e.,

$$|\{n \,:\, \theta_n \in \Theta_0 \setminus B_\varepsilon(\mathcal{A}_\infty)\}| < \infty. \tag{A.17}$$

Otherwise, infinitely many $\theta_n \in \Theta_0 \setminus B_\varepsilon(\mathcal{A}_\infty)$ form a sequence in the compact set $\Theta_0 \setminus B_\varepsilon(\mathcal{A}_\infty)$. This implies that $(\theta_n)_n$ has an accumulation point in $\Theta_0 \setminus B_\varepsilon(\mathcal{A}_\infty)$, which contradicts the definition of $\mathcal{A}_\infty$.

By assumption, $\mathcal{A}_\infty$ is a finite set, and thus, the minimum Euclidian distance between two different points in $\mathcal{A}_\infty$ is positive, i.e.,

$$\delta = \min\{\|\theta^* - \tilde{\theta}^*\| \,:\, \theta^* \neq \tilde{\theta}^* \in \mathcal{A}_\infty\} > 0. \tag{A.18}$$

By (A.14),

$$S(\theta^*) = T_k \cdots T_1(\theta^*) = \theta^*, \quad \forall \theta^* \in \mathcal{A}_\infty. \tag{A.19}$$

Each partial maximization operator T_i is assumed to be continuous, which implies that the operator S defined in (A.19) is also continuous. Then since S is continuous and $\mathcal{A}_\infty$ is finite, we can choose $\varepsilon \in (0, \delta/2)$ such that for all $\theta^* \in \mathcal{A}_\infty$,

$$\theta \in B_\varepsilon(\theta^*) \Longrightarrow S(\theta) \in B_{\delta/2}(\theta^*). \tag{A.20}$$

Note that $\varepsilon < \delta/2$ implies that all $B_\varepsilon(\theta^*)$, $\theta^* \in \mathcal{A}_\infty$, are pairwise disjoint and each ball $B_\varepsilon(\theta^*)$ contains no accumulation point of $(\theta_n)_n$ other than θ^*.

Now it follows from (A.20) that

$$\theta_n \in B_\varepsilon(\theta^*) \Longrightarrow \theta_{n+1} = S(\theta_n) \in B_{\delta/2}(\theta^*), \quad \forall n. \tag{A.21}$$

By (A.17), we can choose $N \in \mathbb{N}$ such that

$$\theta_n \in B_\varepsilon(\mathcal{A}_\infty), \quad \forall n \geq N. \tag{A.22}$$

By definition of δ and ε,

$$B_{\delta/2}(\theta^*) \cap B_\varepsilon(\mathcal{A}_\infty) = B_\varepsilon(\theta^*). \tag{A.23}$$

Thus, by (A.21) and (A.22)

$$\theta_n \in B_\varepsilon(\theta^*) \Longrightarrow \theta_{n+1} \in B_{\delta/2}(\theta^*) \cap B_\varepsilon(\mathcal{A}_\infty) = B_\varepsilon(\theta^*), \quad \forall n \geq N. \tag{A.24}$$

Finally, let $\theta_\infty \in \mathcal{A}_\infty$ be such that $\theta_N \in B_\varepsilon(\theta_\infty)$. Then by (A.24),

$$\theta_n \in B_\varepsilon(\theta_\infty), \quad \forall n \geq N. \tag{A.25}$$

Since θ_∞ is the only accumulation point of $(\theta_n)_n$ in the closure of $B_\varepsilon(\theta_\infty)$, we obtain that

$$\lim_{n\to\infty} \theta_n = \theta_\infty. \tag{A.26}$$

□

Corollary A.5. *If the likelihood equations have only finitely many solutions that lie in the same contour of the likelihood, i.e.*

$$\left| \left\{ \theta \in \Theta : \left. \frac{\partial \ell(\theta)}{\partial \theta} \right|_{\theta=\theta^*} = 0, \quad \ell(\theta) = L \right\} \right| < \infty, \quad \forall L \in \mathbb{R}, \tag{A.27}$$

then the sequence $(\theta_n)_n$ converges.

Proof. Follows immediately from Proposition A.4 since, by Lemma A.2, $\mathcal{A}_\infty$ is a subset of a set of solutions to the likelihood equations that lie on the same contour of the likelihood. □

Proposition A.6. *If $(\theta_n)_n$ does not converge, then the set $\mathcal{A}_\infty$ of its accumulation points is an infinite, compact and dense subset of Θ_0.*

Proof. Non-convergence of $(\theta_n)_n$ is equivalently expressed as $|\mathcal{A}_\infty| \geq 2$. Straightforward extension of the reasoning in the proof of Proposition A.4 implies that if $|\mathcal{A}_\infty| \geq 2$, then there cannot exist an accumulation point $\theta^* \in \mathcal{A}_\infty$ such that

$$\inf\left\{\|\tilde{\theta}^* - \theta^*\| : \tilde{\theta}^* \in \mathcal{A}_\infty \setminus \{\theta^*\}\right\} > 0. \tag{A.28}$$

Thus, $\mathcal{A}_\infty$ must be infinite and the infimum in (A.28) must be zero for all $\theta^* \in \mathcal{A}_\infty$. Therefore, for any $\theta^* \in \mathcal{A}_\infty$ there must exist a sequence in $\mathcal{A}_\infty \setminus \{\theta^*\}$ that converges to θ^*, which makes $\mathcal{A}_\infty$ a dense set.

To show that $\mathcal{A}_\infty$ is compact, it is enough to show that $\mathcal{A}_\infty$ is closed, since as a subset of Θ_0 it is automatically bounded. Let $(\theta_n^*)_n$ be a sequence in $\mathcal{A}_\infty$ that converges to $\theta_\infty^* \in \Theta_0$. Then for any $\varepsilon > 0$, there exists $n(\varepsilon) \in \mathbb{N}$ such that

$$\theta_{n(\varepsilon)}^* \in B_\varepsilon(\theta_\infty^*). \tag{A.29}$$

Since $\theta_{n(\varepsilon)}^*$ is an accumulation point of $(\theta_n)_n$, we can choose $\tilde{\varepsilon} > 0$ such that

$$B_{\tilde{\varepsilon}}(\theta_{n(\varepsilon)}^*) \subseteq B_\varepsilon(\theta_\infty^*) \tag{A.30}$$

and that there exists

$$\theta_{n(\tilde{\varepsilon})} \in B_{\tilde{\varepsilon}}(\theta_{n(\varepsilon)}^*) \subseteq B_\varepsilon(\theta_\infty^*). \tag{A.31}$$

Since ε was arbitrary, θ_∞^* must also be an accumulation point of $(\theta_n)_n$, i.e. $\theta_\infty^* \in \mathcal{A}_\infty$, which establishes the closedness of $\mathcal{A}_\infty$. □

Remark A.7. In this appendix, we derived convergence properties for iterative partial maximization algorithms without relying on any published results. Instead, we could have cited, for example, Wu (1983) and showed that $\|\theta_{n+1} - \theta_n\| \to 0$ as $n \to \infty$. Since Wu (1983) phrases his work in terms of EM-algorithms, we preferred to derive our results "from scratch." Finally, an example illustrating the result in Proposition A.6 can be found in Boyles (1983).

VITA

Mathias Drton was born on May 24, 1975, in Hirschau, Germany. In March 2000, he received a Diplom in "Wirtschaftsmathematik" (Applied Mathematics) at the Universität Augsburg in Germany. During the school year 1998/99, he visited the Université Paul Sabatier in Toulouse, France, where he obtained a Diplôme d'Études Approfondies (DEA) in Applied Mathematics. From March 2000 to March 2001 he worked as a research assistant at the Universität Augsburg. In March 2001, he commenced studies at the University of Washington, and in 2004 he graduated with a Doctor of Philosophy in Statistics.